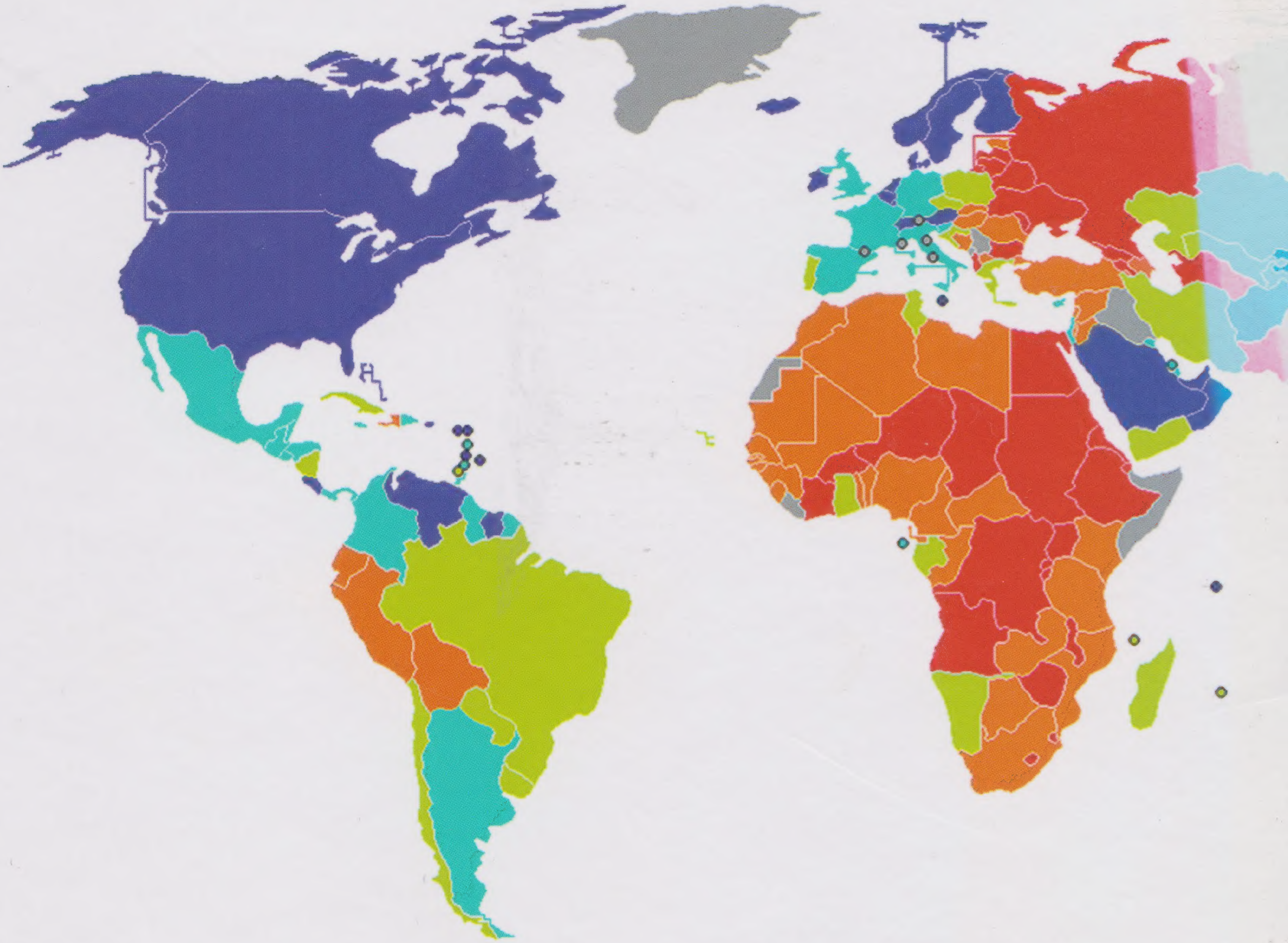
Absolute Certainty and Other Fiction

The secrets of statistics



Everything is mathematical

Absolute Certainty and Other Fiction

Absolute Certainty and Other Fiction

The secrets of statistics

Pere Grima

Everything is mathematical

For Alícia and Pau

Published by RBA Coleccionables, S.A.
c/o Hothouse Developments Ltd
91 Brick Lane, London, E1 6QL

Localisation: Windmill Books Ltd.
Photographic credits: age fotostock.

ISSN: 2050-649X

Printed in Spain

Contents

Preface

We have all heard of statistics. We are accustomed to hearing them being discussed in the media: a study (statistical, obviously) claims that the consumption of drugs has decreased among young people...The results of a survey confirm that the prime minister is more popular than a rival... If an election was called today, one party would win with such and such a majority... Even football commentators discuss statistics, such as one team on the field always scores more goals in the second half. Curiously, although we use the word statistics to refer to the general discipline, actual 'statistics' are nothing more than figures. The information we can deduce from the data and how reliable it is (yet another 'statistic') are not always clear.

Sometimes statistics is regarded as lacking in rigour. Claiming something is not to say it will happen and it is highly possible that on that occasion, the football team who always scores in the second half will not score a single goal in either. Contrast this with mathematics, which has a much more serious image. If a team is the 'mathematical' champion, it will win regardless of what happens. This image of something that lacks solidity is not helped by the perplexity created by a politician's ability to present data and statistics in a way that is always in favour of their theories or interests.

However, there is much more to statistics than this. Statistics is seen in many areas: In medical research (is a new drug better?); in biology (how many of a certain species are there in an area and are they at risk of extinction?); in forecasting (how much electricity will we use tomorrow?); in market research (what type of packaging does the consumer like best?); in sociological studies (what do young people think about an issue?); in economics (how much have prices gone up?); or in quality assurance (which is the problem that needs most attention?). Perhaps this list seems too long, but it is not an exhaustive one by any means. There are many areas in which statistics is of fundamental importance.

Statistics studies how to gather data – how much to gather and in what way – and how to analyse it to obtain information that makes it possible to answer the questions we have posed. It involves making advances in knowledge based on intelligent and objective observations and an analysis of the real world. This is the essence of the scientific method.

This book takes a look at some of the most interesting aspects of statistics, from how to present information using graphs and how to avoid letting in goals – to continue the footballing theme – when they are presented to us, to how to organise gathering data to answer the questions that need asking, such as surveys and election polling, and a standard way of reasoning for all statistical tests. It also considers the calculation of probabilities, an aspect that will perhaps seem dry and difficult to many readers, but which, without the need to go into great depth, reveals many interesting hidden features.

The aim is that this book is both entertaining and educational. If I have managed to achieve this aim, it is thanks to what I have learned from my colleagues at the Universidad Politécnica de Cataluña and from lecturers who are passionate about teaching statistics, such as Roberto Behar at the Universidad delValle in Cali, Colombia. Finally, I would like to express my gratitude to Pedro Delicado, Lluis Marco, Lourdes Rodero and Xavier Tort-Martorell for their detailed reading of the first draft of this book and their wise comments and suggestions which made considerable improvements.

Chapter 1

Descriptive Statistics: How to Get Information out of a Jumble of Data

What should we do when we have a large set of data from which we want to extract specific information? Undoubtedly, the first action is to 'take a look at them', but not looking at them as they are, one after the other (our mind is not great at uncovering information in this way), but through graphical representations or by summarising the data into a few values that can be interpreted more easily.

Historical prelude: the great cholera epidemic of 1854

Today, London's Soho is one of the most alluring districts in the city. Its irresistible mix of modernity and tradition puts it on the sight-seeing list for the many tourists who, year after year, frequent its fashionable bars and eateries and rest their weary feet in the enchanting green squares that emerge here and there between the narrow alleyways. With so many attractions and the characteristic hustle and bustle of all large city centres it is quite unlikely that an exact replica of a 19th-century water fountain in a corner of Broadwick street will be noticed. However, this modest monument, and it is a monument, commemorates an event of such importance that it could be raised to a height of one hundred metres and illuminate the London sky with a spotlight.

Erected in 1992 in honour of British epidemiologist John Snow, the fountain on Broadwick Street (formerly known as Broad Street) is located just a few metres from another identical one which, in 1854, pumped water from the Thames for use in the neighbourhood. In August of that fateful year, a brutal cholera epidemic was declared in the area, killing 100 people in three days and, after two weeks, the toll was 500. More than three quarters of the population abandoned their houses to escape the foul smell that was thought at the time to be the way this terrible disease was spread.

John Snow, an eminent doctor, who one year earlier had personally administered chloroform to Queen Victoria while she was giving birth for the seventh time, disagreed. In a text written in 1849 he argued that the cholera was not being transmitted through the air but in the water. The medical community paid little attention to his opinion, largely because it was not backed up by any specific theory on exactly what the water contained to cause the sickness. Snow's convictions were based on a veritable arsenal of observations in which an unavoidable connection between the water and the transmission of cholera was established. It was 'merely' statistical evidence, of a relation between a cause and effect for which, as it has been said, Snow *had no explanation*. Despite this, Snow's observations were so convincing, and his ability to demonstrate them so great, that his contemporaries had no choice but to accept his opinion and, in doing so, the way in which modern cities provide water to their population was transformed forever.

Hunting down the culprit

Cholera is a terrible illness the main symptoms of which are sudden and intense vomiting and diarrhoea that can cause fatal dehydration. The cholera outbreak of 31 August 1854 was very quickly recognised as "the worst in the country's history". The figures are horrifying: in 72 hours the number of victims had already reached 127, many of them children. Three days after the outbreak, Snow visited the area in the company of local reverend Henry Whitehead, and he discovered that most of the deaths had occurred in houses near the public water fountain on Broad Street where it intersects Cambridge Street. Snow noted:

> "On proceeding to the spot, I found that nearly all of the deaths had taken place within a short distance of the pump on Broad Street. There were only ten deaths in houses situated decidedly nearer to another street pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pump which was nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street."

When he examined the pump he found no visible sign of contamination. Next, he checked the forensic records and made a detailed list of the deaths from the past two days. None of the employees of a brewery near the pump had contracted the disease, and a homeless shelter also located nearby, which served more than 500 people had recorded a very modest 5 cases.The daily reports on the epidemic spoke of new victims further afield in areas such as Hampstead and Islington. It seems like Snow's theory was crumbling.

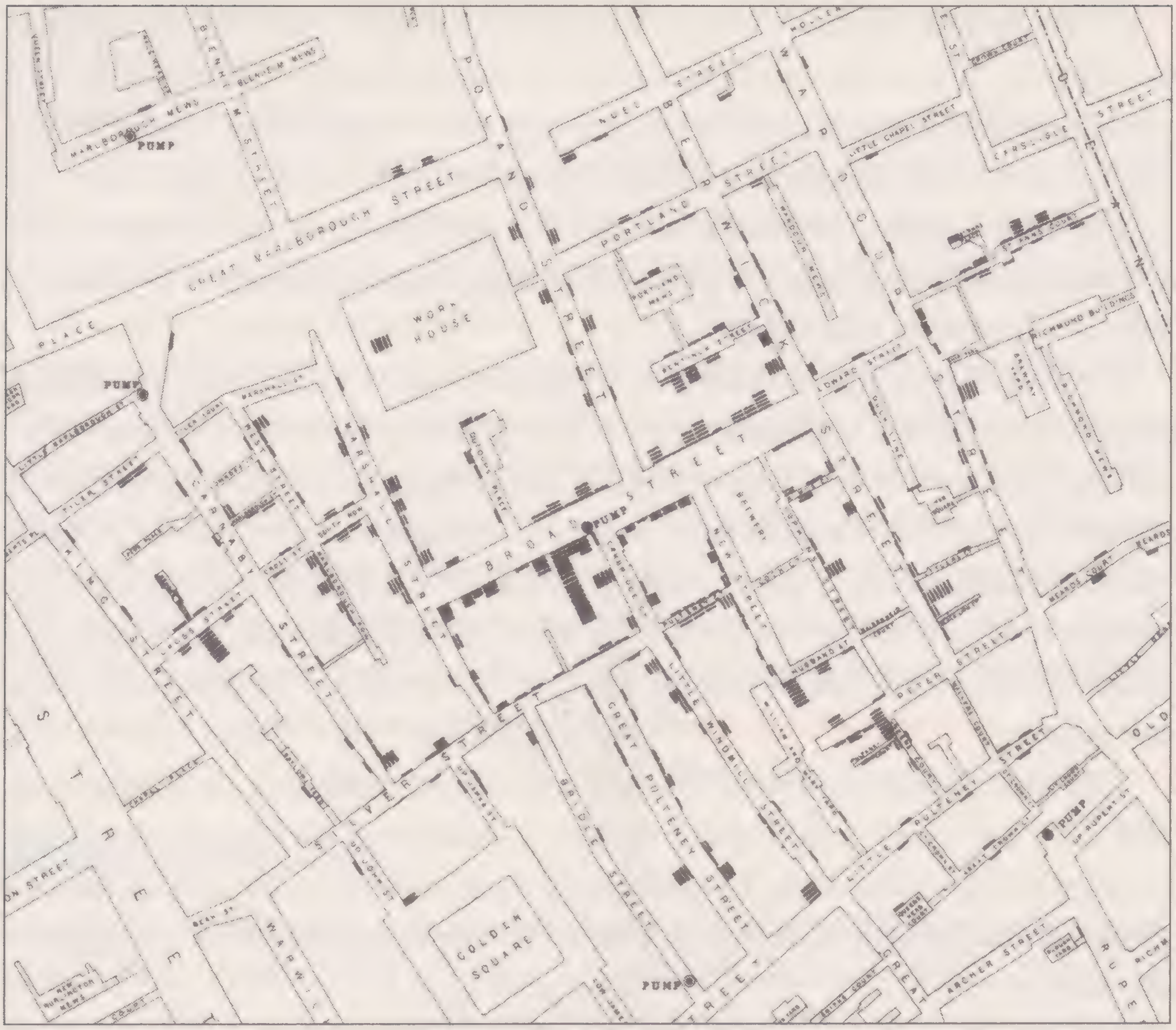
However, the doctor intensified his efforts. He went from building to building, from house to house, and found out that both the homeless shelter and the brewery had their own wells for water and did not use the pump. Once in Hampstead, the family told him that the victim, a woman, brought a bottle of water from the water fountain on Broad Street every day because "she liked its taste better".The woman's niece, who had also recently died of cholera, used to do the same. "And where did she live?", we can almost imagine Snow asking. "In Islington," came the answer.

The doctor modestly wrote:"The conclusion of my investigation is, consequently, that there is no cholera outbreak or significant presence of the illness in this part of London except among those who habitually drink water from the aforementioned pump."A brief paragraph, but one that would revolutionise public health throughout the world.

On 7 September, with the epidemic still in full sway, Snow called an urgent meeting with the local authorities and informed them of his findings.As well as the oral report, Snow showed a map of the area on which he had marked the number and location of the victims.The map was so convincing that the next day the pump's handle was removed. The number of deaths plummeted and, in a short time the epidemic was completely over.

The power of a graph

Snow's original map is now kept in the British Museum. In 1855 an improved version was included in a revised edition of his text from 1849, a fragment of which is shown overleaf. It is perhaps difficult for the modern reader to get an idea of how revolutionary the way in which Snow presented his information was, as nowadays the use of graphical representation is very common for demonstrating information.



A fragment of the map of Soho where the cholera epidemic broke out in 1854. The pump on Broad Street, labelled "Pump", is in the centre of the map. The horizontal bars represent the victims in each house.

By representing each victim with an individual mark (the parallel bars), giving each one the same thickness and placing them, house by house, on a conventional map, the geographical component of the epidemic became immediately visible. It is obvious that most of the deaths were building up around the pump on Broad Street in the centre of the map. Adding to this the ardent field work carried out by Snow, the idea that the infection of the disease was directly related to the pump did not require any specific theory on the nature of the relationship. Local authorities understood that, and the result of dismantling the pump was not only the timely ending of the epidemic, but also the confirmation that cholera could be transmitted by means of water. The experiments carried out between 1860 and 1864 by Louis Pasteur would be key when it came to consolidating the theory of germs and pathogenic agents, thus supporting Snow's observations. In 1885 the German Robert Koch identified the bacteria *Vibro cholerae* as the cause of the disease, and towards the end of the century most Western cities had renewed their drinking water supply networks, thus exorcising the ghost of cholera from the streets of half the world.

Summarising data: measures of central tendency

If you had to describe the face of a suspect so that others could clearly recognise it you would have difficulties, unless the suspect had a very distinctive feature. However, police experts know which characteristics to focus on and which adjectives to describe them so that someone else who also has a good command of the jargon can also get an idea of what the individual in question looks like and, if necessary, they also know how to draw them so that others can identify them.

In statistics we do something similar. To summarise the information in a large set of data, a few measurements are selected to concentrate the maximum information and with just a glance (at maybe 5 or 6 values) we can get a reliable idea of the general behaviour of the data. These measurements are normally presented divided into three groups: central tendency, dispersion and position. In this section we will describe the first group, which indicates around which values the data is centred.

The arithmetic mean

All of us probably learned to calculate means, or averages, at school. On a scale of 0 to 10, if 5 or more is a pass and the final mark is the mean of three partial exams, receiving 3, 2 and 6 is a fail and marks of 4, 4 and 7 is a pass (what if they are 4, 4.5 and 6.3?).

The arithmetic mean is the measure *par excellence* of central tendency. Its good properties, together with the fact that it is easy to understand and calculate make it very popular. It also has less trivial aspects, such as when carrying out operations with means. The mean of (3, 4, 5) is 4, and that of (4, 6) is 5, but the mean of the set is not the mean of the means $(4+5)/2=4.5$, but 4.4. In general, if we have a set of n_1 values whose mean is $\overline{x}_1$ and another of n_2 values whose mean is $\overline{x}_2$, the mean of the set $\overline{x}_T$ will be:

$$\overline{x}_T = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}.$$

This is exactly the same as takng the mean of all the values, because if a sample has n members and its mean is $\overline{x}$, the sum of all of them will be $n\overline{x}$; thus, in the numerator we have the sum of all the values and in the denominator, the number of values.

Let's look at an example. If the mean age of a company's workers is 36 years for men and 32 for women, what is the global mean? It depends on the proportion of men and women. If half are men and the other half women, the average age will be 34 years. If there are 75% men and 25% women, it will be 35. Note that in the above formula the proportion represented by the first set of data is $p_1 = n_1/(n_1 + n_2)$ and by the second $p_2 = n_2/(n_1 + n_2)$, so it can be written:

$$\overline{x}_T = p_1\overline{x}_1 + p_2\overline{x}_2.$$

There are also cases in which the mean is not the most suitable measurement. If we are trying to summarise the time that a supplier takes to supply a product, or how long a train takes between two towns, the mean is a bad indicator of the quality of service. It could be that the agreed delivery time is 10 days and half the time the material is delivered in 2 (the client does not expect it, there is no room to store it, etc.) and the other half in 18 (the client is now desperate), but on average the supplier is complying perfectly. The same thing happens with the train. Getting to work half an hour early some days (especially if we cannot go in before the office opens) does not make up for the other days when we arrive half an hour late. In these cases, a more informative measurement would be the percentage of times that it arrives late or with more than a specified delay.

Another problem with the mean is that it is highly influenced by extreme values. It would undoubtedly be surprising to learn that most people have an above average number of legs, but it is true, as some only have one or none (in extreme cases), and that makes the average slightly below two.

The median

The median is the value that ends up in the centre when sorting data from lowest to highest. If the data is 6, 7, 5, 2 and 9, the median is 6, which ends up in the centre once they have been sorted into ascending order. If there is an even number of values, nothing will be in the centre, and in this case the median is the average of the two central values. The median's properties cover some of the mean's weak points. So it is more robust in the presence of anomalies. Here is a simple example of why. In the data above the mean is 5.8 and the median is 6. If we make a mistake while entering the data on the keypad and type 99 instead of 9, the average changes to 23.8, while the median continues to be 6. When working with data that is yet to be refined, using the median can be more effective than the mean because the information provided is less affected by possible anomalies.

Another advantage the median has over the mean is that the median always leaves 50% of the observations above and 50% below. For example, if we want to know if we are among those whose earn the most in our company, we have to compare our salary with the median, and not the mean. If there are 10 workers and their monthly salaries are (in thousands of pounds): 0.8, 0.8, 0.9, 0.9, 1.0, 1.0, 1.1, 1.1, 1.2 and 10, all but 1 (in this scenario that's 90%) are below the average, which is 1.88. This never happens with the median. If the our salary is above the median, we are among the 50% who earn the most.

Another example: if the pass mark for an exam is greater than or equal to 5 and the average mark given to the students is 5, we do not know how many have passed. If 50 students took the exam, it could be that 41 failed it with a score of 4; 8 got a 10 and 1 was given a 6. This gives a mean of 5, although these are very unusual results. By contrast, if the median is 5, it is certain that half have passed.

The mode

When talking about measures of central tendency, the mode is also always mentioned. It is the value that is repeated the most. If the values are 0, 2, 7, 2, 8, 2, 5, 4, the mode is 2. The benefits of the mode makes more sense with qualitative data. Thus, for example, in a sample of newly born children it is noticed that the most common eye colour is brown, it would be said that the mode of eye colour is brown.

Summarising data: measures of dispersion

You may have had this rather average discussion: if someone eats a chicken and someone else does not eat any, statistics states that on average they have eaten half a chicken each. Or that if you go to the kitchen and put your feet in the fridge and your head in the oven, your body will have the perfect average temperature. The problem lies in trying to summarise the information only with averages and without

FLORENCE NIGHTINGALE

In the summer of 1853, after completely destroying the Turkish army, the Russian fleet in the Black Sea was ready to take Istanbul and control the Bosphorus strait, threatening both British communications with India and French interests in the Mediterranean. Thus, Britain declared war on Russia and sent troops, who were joined by the French and the Turks, to the Crimean peninsula. It was the start of the Crimean War, which by the time it ended in 1856 had caused thousands of deaths.

Cited as the worst managed of all the wars in which Britain has participated, it is the first of which we have photos and also the first that was followed by newspaper reporters. This may seem like a detail of little importance, but in their reports the journalists explained the terrible living conditions of the soldiers and the disasters due to military incompetence. This created a strong sense of indignation among the public and forced the British Minister of War to send a body of nurses under the management of a devoted, intelligent and practical woman named Florence Nightingale.

When the nurses arrived at the hospital in Istanbul, it was in complete chaos. Florence Nightingale explained that most of the deaths were due to infectious diseases and not the wounds that the soldiers had arrived with. She saw, understood and quantified the association between the overcrowding of patients and the death rate, focussing on improving cleanliness, nutrition and order in caring for the sick.

The fact is that during the first seven months of the war, before Florence Nightingale arrived, a British soldier wounded on the battle field had more chance of surviving if he stayed on the front than if he was transported to a military hospital. However, during the last six months of the war, after the changes introduced in the hospitals, the death rate fell from 40 to 2%.

Florence Nightingale knew how to select the data that showed the reality as it was, and carried out the analysis and comparisons needed to understand what the problem was and what measures needed to be taken. And with that statistical analysis, expertly explained, she was able to take on the bureaucracy and conservatism of the army and convince the military brass of the need for radical change in hospital conditions. She saved many lives, and many of the procedures that she introduced are still standard in today's hospitals. As a result of her work, Florence Nightingale became a household figure in the UK and was the first woman to be accepted into the British Royal Statistical Society.

paying attention to the variability of the data. Another example that demonstrates the same error is identifying the well-being of the citizens of a country by only taking into account the per capita income (that would equate to the half chicken example). If you had been given the choice of which country you wanted to be born in, it would undoubtedly have been useful to check not only the per capita income but also the variability. It is better to be born in a country where everyone has their quarter of a chicken guaranteed than in one when the average is half a chicken but where you would have a good chance of not getting any. In short, to summarise information that contains such data it is also necessary to quantify their dispersion, and to do that there are various measurements at our disposal, as we will see below.

Range

The range is the difference between the maximum and minimum values. For example, if the values are 2, 6, 7, 12, 12, 18, the range is $18 - 2 = 16$. It has the advantage of being a very simple measurement, but the inconvenience is that it does not make use of the information contained in the data. Using only the extremes, which can also be exceptions, is a very poor indicator, especially if the set of data is large. If we have few values (in the order of 4 or 5) the range is not such a bad measurement. If we only have two it is as good as any, but in that case there would be no need to summarise it.

Variance and standard deviation

The most widely used measurement of variability is the standard deviation, but in order to define that, it is best to start with variance, as the standard deviation is simply the square root of the variance.

If we had to design a measurement of dispersion, the first idea would be to include all the values we have, as was the case with the mean. For example, if the values are 1, 2, 4, 7 and 9, we can calculate the average of the difference between each of the values from the mean, which is 4.6:

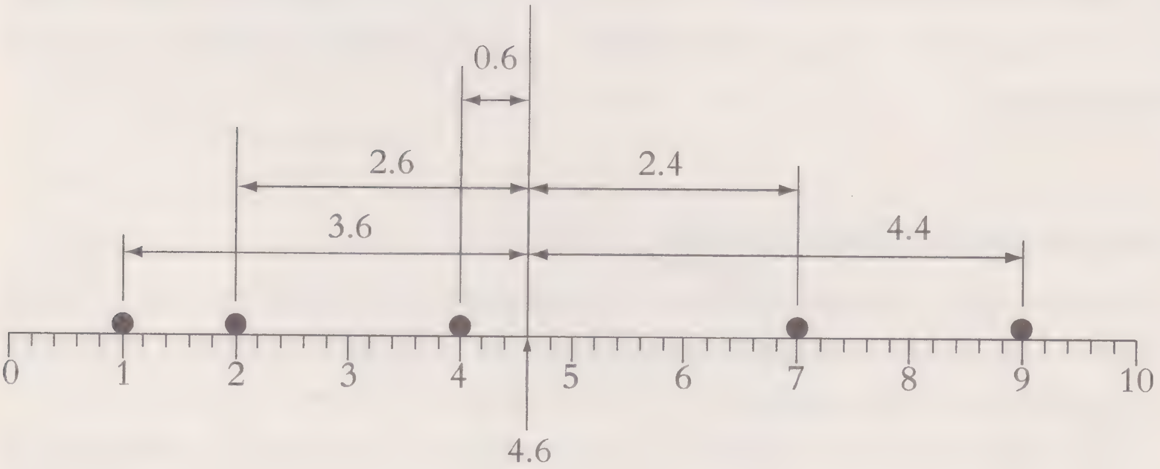
$$\frac{(1-4.6)+(2-4.6)+(4-4.6)+(7-4.6)+(9-4.6)}{5}=0.$$

The problem with this measurement is that it always gives zero, whatever the values included – and therefore it does not measure anything, always giving the same value whether there is a lot or a little dispersion). The most obvious solution is to use the absolute value of the differences:

$$\frac{|1-4.6|+|2-4.6|+|4-4.6|+|7-4.6|+|9-4.6|}{5}=2.72.$$

This measurement is called 'mean deviation' and is a good measurement (the greater the dispersion of the values, the greater the value obtained). However, the value given by resolving the problem of the differences balancing out through squaring them has some much more interesting properties.

$$\frac{(1-4.6)^2+(2-4.6)^2+(4-4.6)^2+(7-4.6)^2+(9-4.6)^2}{5}=9.04.$$



The distance of each value compared to the mean (4.6). The variance is the average of the squares of those distances.

This is what we call variance, and it is not only useful for measuring dispersion, but it can also be found at the heart of most of the theory and of statistical methods. It is represented by σ^2. The inconvenient thing about the variance is that its units are the square of those of the data. When dealing with lengths measured in metres, the units of the variance are metres squared and this complicates the interpretation a little. The solution is very simple – we square root it.

A COUPLE OF FORMULAE

The general formula for variance is:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N},$$

where x_i represents each of the values; μ, their average, and N is the number of values. The corresponding formula for standard deviation is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}.$$

This result, represented by σ, is called the standard deviation, and is the best measurement of dispersion. The standard deviation is often paired with the mean in order to summarise the characteristics of the data with just two values.

Coefficient of variation

What has greater variability: the weight of cats or the weight of cows? Let's assume that the mean weight of cats is 4 kg and that in 95% of cases it is between 3kg and 5kg, and that a certain breed of cow has a weight between 480kg and 500kg, also in 95% of cases. If you analysed a group of cats you would find a great deal of variance among them (some weigh nearly twice as much as others), while cows almost all look the same.

However, the standard deviation of the weight of the cats will be around 0.5kg (according to the weight variability patterns, 95% of the individuals are in the mean interval ± two standard deviations, which will be looked at in more detail in the next chapter). Meanwhile in the cows the standard deviation is 5kg, 10 times more, but with less variability. To resolve this paradox, which arises in the comparison of variabilities, a coefficient of variation is used. This value is the quotient between the standard deviation and the mean:

$$CV = \frac{s}{\overline{x}}.$$

This is called normalising the variability with respect to the mean. In our example, we get 0.125 for the cats and 0.01 for the cows, without units, as it is a non-dimensional measurement.

TWO WAYS TO CALCULATE THE STANDARD DEVIATION

The variance and the standard deviation, despite being at the centre of statistical theory, have a problem that is often kept secret. When we are trying to summarise the information contained in a set of data, we can find ourselves in one of the following situations:

1. The data we have is the object of our interest. We want to know the average or the standard deviation of that data, which constitutes what we call 'population'.
2. The data we have is a sample of the population being studied. In other words, what we are interested in is not so much finding the average or the standard deviation of the data we are dealing with as estimating values for the population.

For the average it does not matter which situation we find ourselves in. The formula is the same, as the best estimate for the average of the population is the average of a sample. It is necessary, as always when we want to reach conclusions on the population through a sample, that the sample is representative.

In the case of variance, things change slightly. If we have the population, the formula we use is the one deduced previously, but if we have a sample – and the aim is to estimate the variance of the population – the formula we use is the following:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}.$$

Why? The problem resides in the fact that when working with samples, the variability is calculated around the average of the sample itself (not around the average of the population, which is what we are actually interested in). It could be said that the average of the sample corresponds to the

Summarising data: measures of position

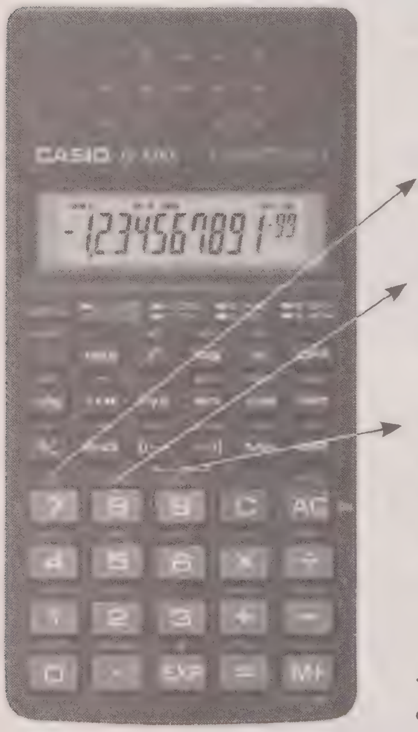
There are measurements in common usage that are not of central tendencies or dispersion. They are used for establishing milestones, delimiting areas in the data set by being reference points for locating values.

Quartiles

With the data sorted from lowest to highest, the median is the point at which the data divides it into two halves. The first quartile is the median of the first half, therefore

data in the sample itself and this tends to undervalue the variability of the population. Dividing by $n-1$ gives a slightly higher value which better estimates the variance of the population. Dividing by 4 is different from dividing by 3, but there is not much difference between dividing by 100 and by 99. For practical purposes, when the sample is large, this issue ceases to be relevant.

Do not worry if this seems like a real muddle and you did not understand it. Carry on regardless and if at some point you have to choose, work as if you have a sample (use the formula that divides by $n-1$). If you use statistical software that does not give you the option, this is what it will be doing.



$\overline{x}$:Average of the data.

σ_n: Standard deviation of the data when dealing with the entire population (the standard deviation of the relevant data).

σ_{n-1}: Standard deviation of the data when dealing with a sample and attempting to estimate the standard deviation of the population from which the sample comes.

Statistical functions on a calculator: one key for the average and two for the standard deviation.

leaving 25% of values below and 75% above. The median of the second half is the third quartile, which leaves 75% below and 25% above.

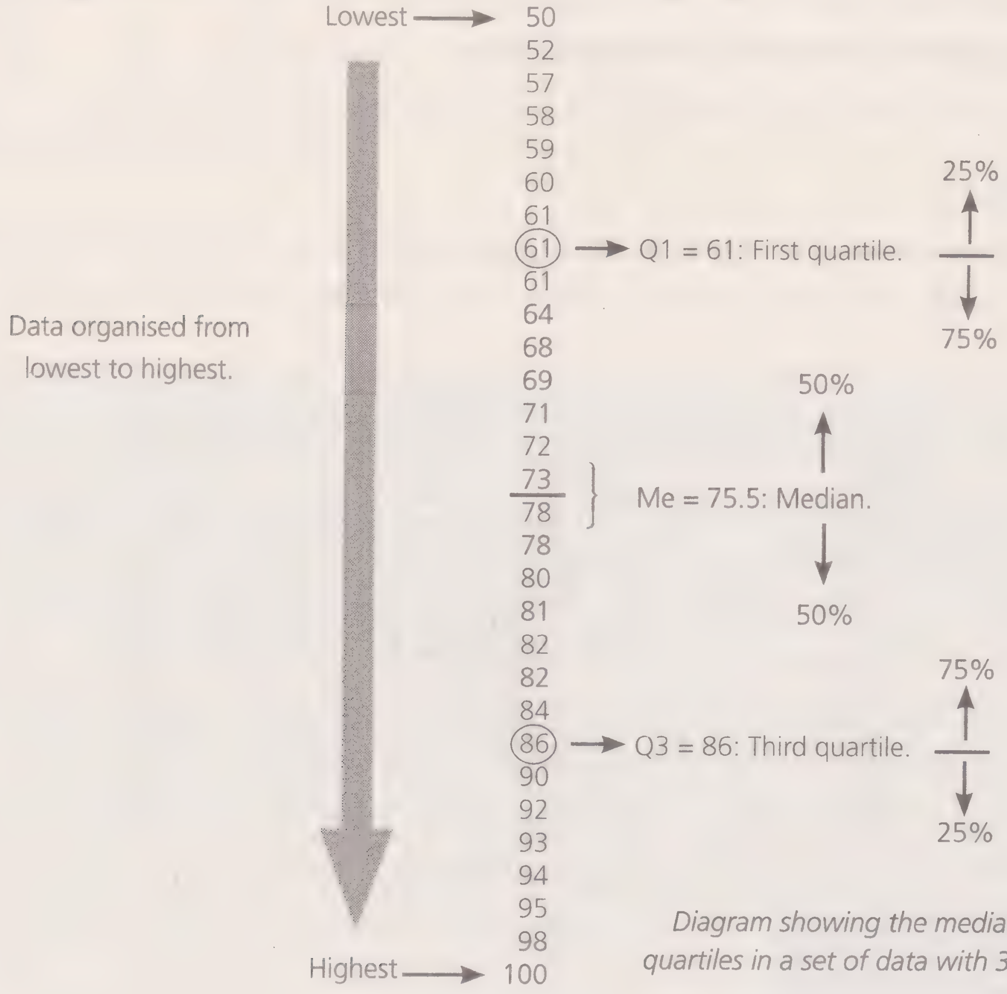


Diagram showing the median and the quartiles in a set of data with 30 entries.

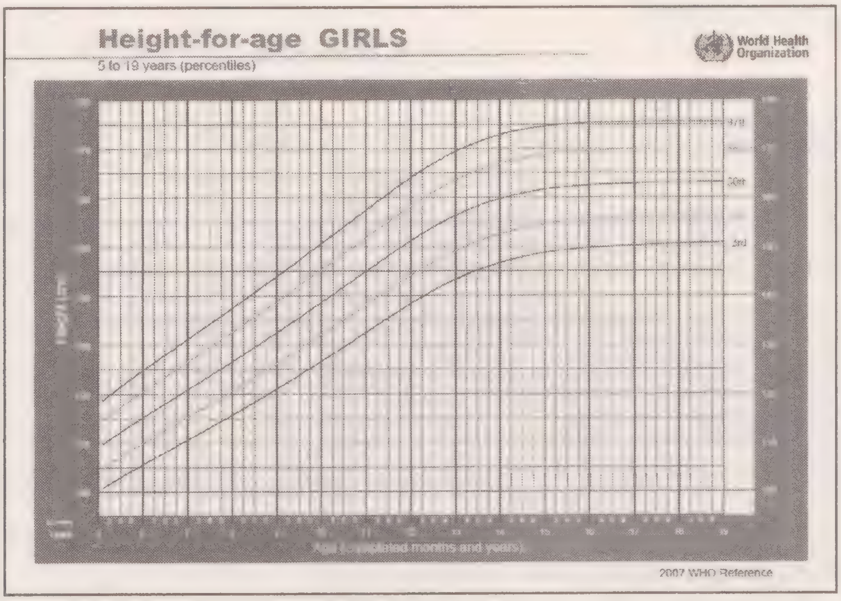
Thus, if in a company the first quartile of monthly salaries is £1,000 and the third quartile is £2,000 then your pay of £800 places you in the 25% who earn the least. If your salary is £1,500 you are in the half who earn the most, but at least 25% of the employees earn more money than you. If you earn £2,100 a month, you are among the privileged 25% who earn the most.

Percentiles

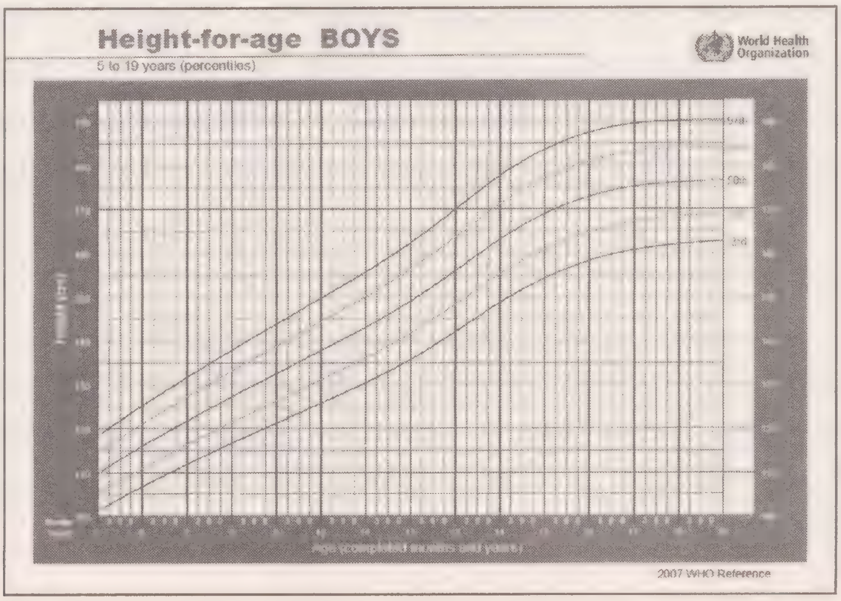
The 15th percentile is the value that – with all the data sorted – leaves 15% of it below and, therefore, 85% above. The quartiles are the 25th and 75th percentile and the median is the 50th.

Going back to the example of the salaries, if yours is in the 70th percentile it means that 70% earn less than you (or that 30% earn more, if you are a glass-half-empty person). Percentiles are also used to measure the results of aptitude tests. If you are in the 90th percentile it means that 90% of the population who take the test are worse at the ability being measured.

Many people have their first contact with percentiles when their child's midwife says that the child's height is in the 45th percentile. This means that 45% of children (the references are different for boys and girls) of their age are shorter than they are. The World Health Organisation publishes reference tables and graphs on the growth of different age groups.



Reference graphs published by the World Health Organisation with the median and the 3rd, 15th, 85th and 97th percentiles for the statures of children from 5 to 19 years old.



Percentages: they seem harmless but they are dangerous

One way of highlighting a relevant aspect of a set of data is to use percentages ("65% of minors from 10 to 17 years old admit that they play video games designed for adults"), but statistics books do not generally deal with the subject, perhaps because they think it does not come under their jurisdiction or because they consider it too easy. The percentage symbol on the simplest calculators is next to the add, subtract, multiply and divide symbols, so it may seem like it is a subject that anyone with a basic knowledge of arithmetic would have a good command of. However, percentages can cause confusion and misunderstandings, therefore it is well worth spending a little time on this subject.

General problems

We should always keep in mind what percentage is. Let's take a look at an example: A gel is normally sold in 750 ml bottles and now, for the same price, it is a 1 litre bottle. What percentage of gel are they giving away? It depends on which value the percentage is calculated. If we use the first value, then 33% is being given away, and with the second value it is 25%.

We should also distinguish between percentage and percentage points. Thus, if we say that a company's profits have risen from 2 to 4% then they have increased by two percentage points (but not by 2%!).

In the same way, we should distinguish between percentages based on levels and percentages based on changes in level. The following example explains this problem. Last year a salesman sold £10 million-worth of goods. His target for this year was to increase that number by 6%. The salesman managed to sell just £10.3 million worth. What percentage of the target has he achieved? If his goal was making an increase, he only achieved 50%, but if we interpret his goal as hitting a sales target of £10.6 million worth and he sold £10.3 million, he has achieved 97.2% of the target.

Lastly, care must also be taken when doing calculations with percentages:

1. If the price of a product increases by 20% and then decreases by another 20%, how does the final price compare to the start price? The two will not be the same, but have decreased by 4%. If the initial price was X, the final price will be $(X + 0.2X) - 0.2(X + 0.2X) = X - 0.04X$.

2. A product is comprised of 10 components and each component increases its cost by 2%. By how much does the cost of the product increase? It increases by 2%. It does not matter if there are some very cheap components and some very expensive ones. If you don't believe it, do the sums.

3. If John earns 1,000% more than Peter, he earns 11 times more (not 10). If he earns 100% more he earns twice as much, if he earns 200% more, this is three times as much, etc.

It is not what it seems: Simpson's paradox

If global percentages are given comparing groups that in turn contain various parts, they can give the impression that one thing appears to be happening, when in reality something else is going on. This phenomenon is known as the Simpson paradox. Let us consider an example.

A large company opens a new factory creating 250 jobs in the buying, installation and storage departments. In total 355 men and 325 women apply, of which 190 men (53.5%) and 60 women (18.5%) are accepted. It is verified that the level of ability of the men and women is similar in each category. Do the headline percentages show that the women have been discriminated against? The answer is no. The data is as follows:

Department	Places	Applicants		Accepted		% Accepted	
		Men	Women	Men	Women	Men	Women
Buying	30	25	100	5	25	20	25
Assembly	200	250	25	180	20	72	80
Storage	20	80	200	5	15	6.25	7.5
TOTAL	250	355	325	190	60	53.5	18.5

Actually, in all the departments the proportion accepted was greater among the women. The key lies in the fact that a lot of men and few women applied to the department that offers the most jobs, while the opposite happened in those with fewer positions.

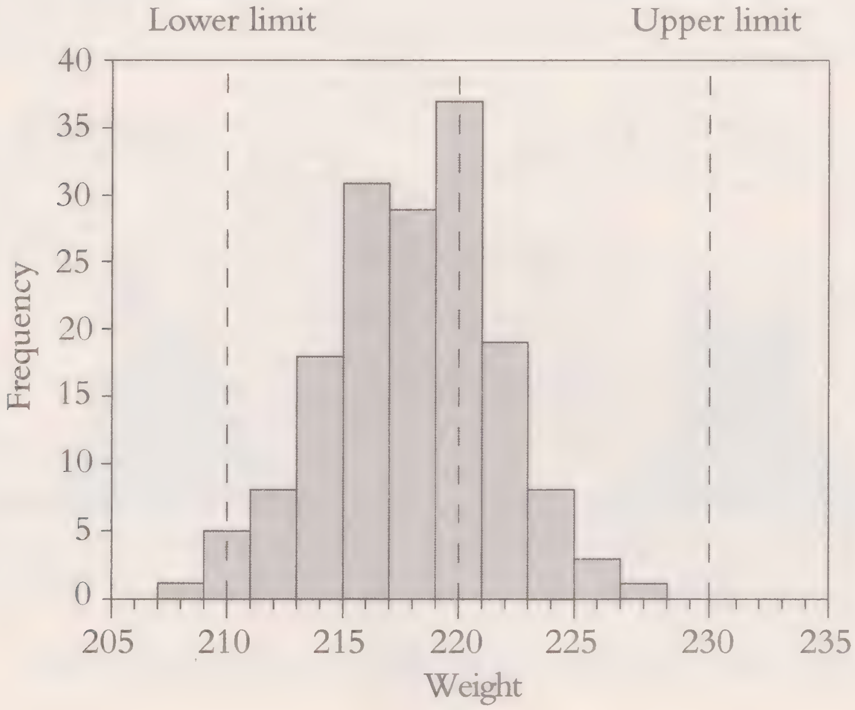
Graphical representations of a variable

Let's start by setting a challenge: the owner of a bakery is worried because he suspects that the weight of the loaves of bread he sells is too variable, meaning that some could even be below the limits set out by law. Two machines are used to make the bread and there are also two operators working there, some days one of them makes the bread and the other days the other does it. The following table contains the weights (in grams) of samples of bread collected at random over the last 20 days.

Day	Operator	Machine 1				Machine 2			
1	A	220.3	215.5	219.1	219.2	220.3	208.0	214.4	219.2
2	B	215.8	222.0	218.9	213.6	216.9	213.4	217.7	217.7
3	B	220.4	218.7	218.6	219.6	222.9	219.7	209.4	221.6
4	B	221.5	227.0	219.5	222.5	223.1	215.3	220.4	215.6
5	A	215.7	225.3	223.0	218.0	216.0	210.9	221.4	210.9
6	A	222.7	215.1	219.6	217.3	212.1	213.0	218.0	216.5
7	A	216.0	218.8	217.9	213.0	216.9	216.0	213.5	219.2
8	B	219.4	218.3	216.7	224.1	216.2	218.4	216.6	214.9
9	B	219.8	222.6	219.1	217.7	216.2	212.2	216.9	214.9
10	A	220.2	219.5	222.4	219.9	222.9	214.3	219.1	216.7
11	B	218.0	223.9	219.6	221.9	214.9	212.6	219.4	213.3
12	B	219.3	219.6	218.8	219.9	219.0	216.7	216.4	213.5
13	B	220.0	214.1	224.3	217.4	218.0	219.5	219.5	222.3
14	A	223.9	220.6	219.5	219.6	211.8	218.2	218.3	217.4
15	A	218.1	218.8	218.4	217.9	214.6	215.7	218.0	216.4
16	B	216.9	221.6	220.6	222.6	215.6	220.4	217.3	216.2
17	B	217.9	225.7	222.2	216.1	212.5	214.6	209.7	211.3
18	A	224.2	216.2	219.9	220.4	215.8	219.9	216.5	211.9
19	A	214.1	219.7	222.4	224.5	213.7	209.7	216.9	213.1
20	A	221.1	225.0	222.7	222.2	212.5	217.5	217.4	215.7

The weight must be 220 ± 10 grams, and we assume that this data is representative of general production. The questions we want to answer are: is there a problem? If so, what is happening? What needs to be done to solve the problem, if there is one?

If you try to draw conclusions by simply looking at the data it is possible to make a mistake. Although in this case there are 160 values, trying to draw conclusions 'by eye' is always risky. Nor is it necessary to do big calculations or apply sophisticated techniques; it is sufficient to represent the data on a graph, as seen below.

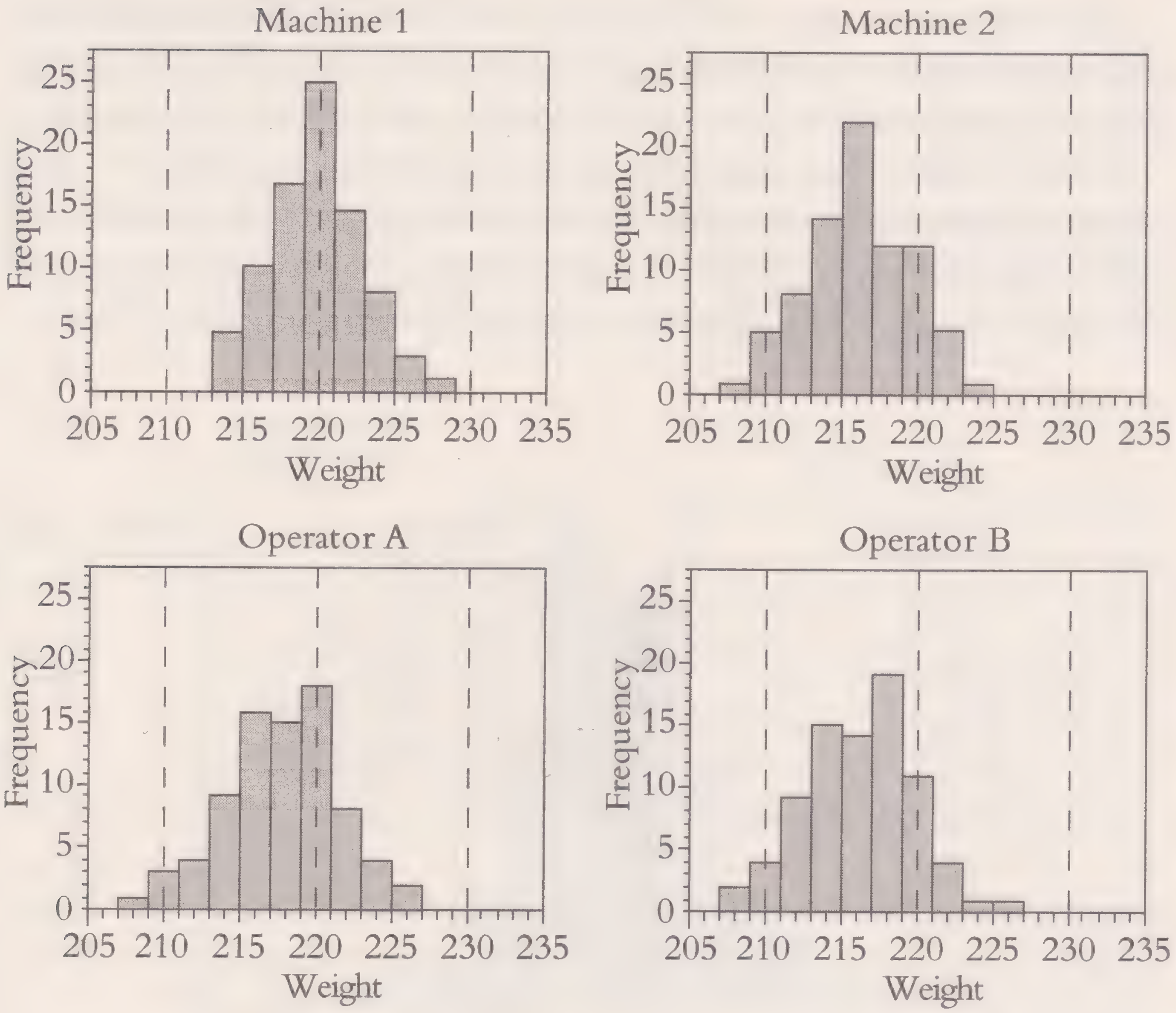


A histogram of the weights of 160 loaves of bread.

This graph is called a histogram and it is very useful for analysing variability in data.

In our example of variability of the weight of loaves of bread, the histogram demonstrates that there is a problem, as there are some loaves that are naturally outside the acceptable limits. In other words, they are not exceptions, but form part of the natural variation in the bread-making process.

Sorting by operator and by machine, as shown by the following histograms, it can be seen that the problem is with machine 2, which is offset. There is no problem with machine 1 and the two operators give practically equal results.

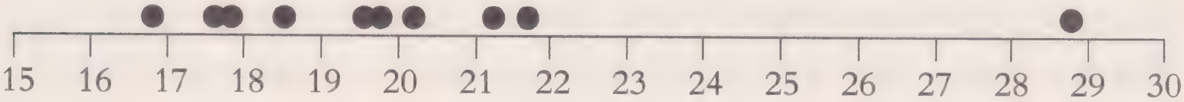


The weights of the loaves of bread sorted by machine and by operator.

Even when using very few values, for example:

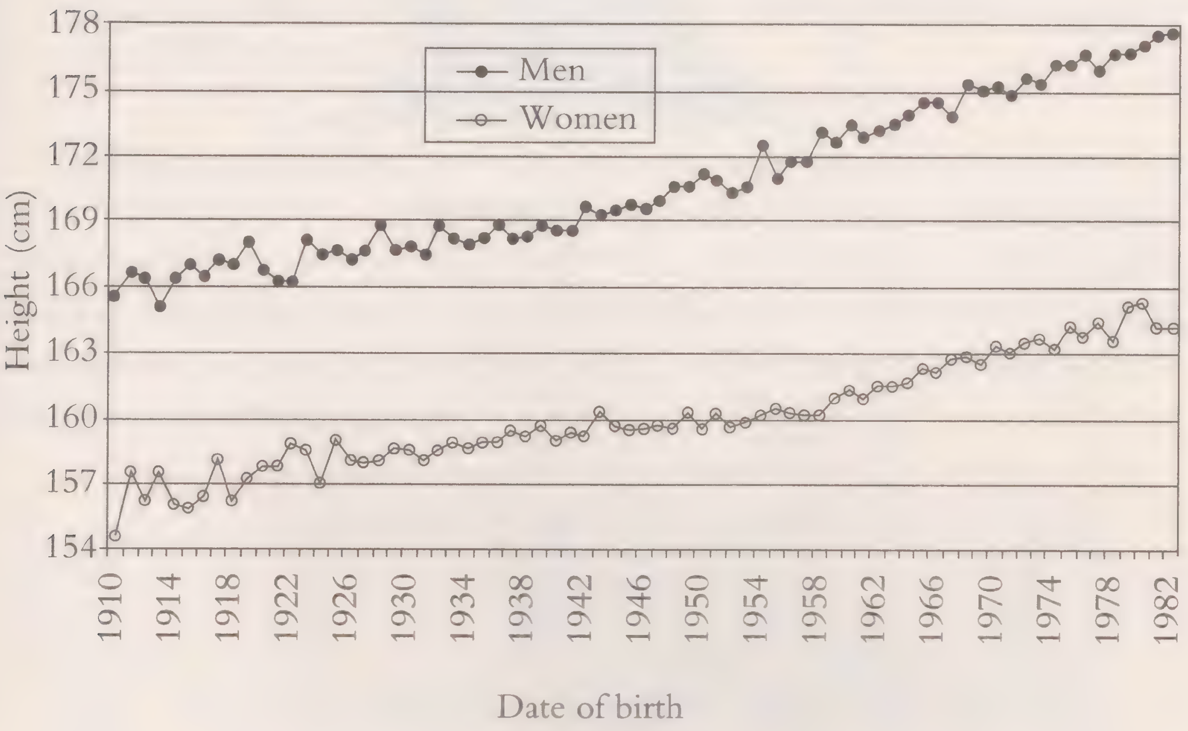
21.1 17.8 19.7 18.6 16.8 21.7 28.7 20.1 19.5 17.8

a simple dot diagram demonstrates the details that can go unnoticed when looking at data alone. In this case there is a value that is significantly separated from the rest and it would be convenient to analyse the cause of this ‘outlier’. Maybe it is just a typing error and it should be 18.7 instead of 28.7. These issues are important, as working with erroneous data can distort the results of the study.



Dot diagram of a set of data.

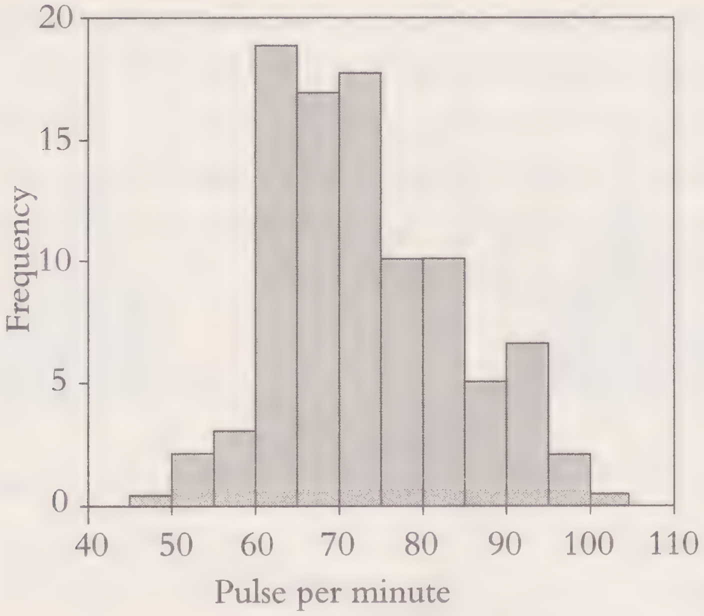
When we want to take into account the order in which the data has been taken, histograms and point diagrams are of no use. What we do is represent them with a time series dot diagram, such as the one in the figure below, which demonstrates the growth of the mean height of Spaniards throughout much of the 20th century. Of course, extrapolating this graph has little value. It is not at all likely that we will measure 2.7 metres within the next 1,000 years.



Generational evolution of the mean height of the adult population in Spain, 1900-1982. (Source: J. Spijker, J. Pérez and A.D. Cámara: "Generational changes in height in Spain during the 20th Century from the National Health Survery", Spanish Statistics *magazine, no. 169, 2008).*

Histograms are typical graphical tools, as are pie charts and line graphs. However, other less well-know methods can be used, such as stem and leaf diagrams.

Let's have a look at a practical example. In a class of 92, the students were asked to take their pulse for one minute, and the histogram on the next page represents the values obtained. (All the values used in this example form part of one of the files included in Minitab, the statistics software package.)



Stem	Leaves
4	8
5	4 4
5	8 8 8
6	0 0 0 0 1 2 2 2 2 2 2 2 2 2 4 4 4 4
6	6 6 6 6 6 8 8 8 8 8 8 8 8 8 8 8
7	0 0 0 0 0 0 2 2 2 2 2 2 4 4 4 4 4
7	6 6 6 6 6 8 8 8 8 8
8	0 0 0 2 2 2 4 4 4 4
8	6 7 8 8 8
9	0 0 0 0 2 2 4
9	6 6
10	0

Histogram and stem and leaf diagram of the pulses per minute of 92 students.

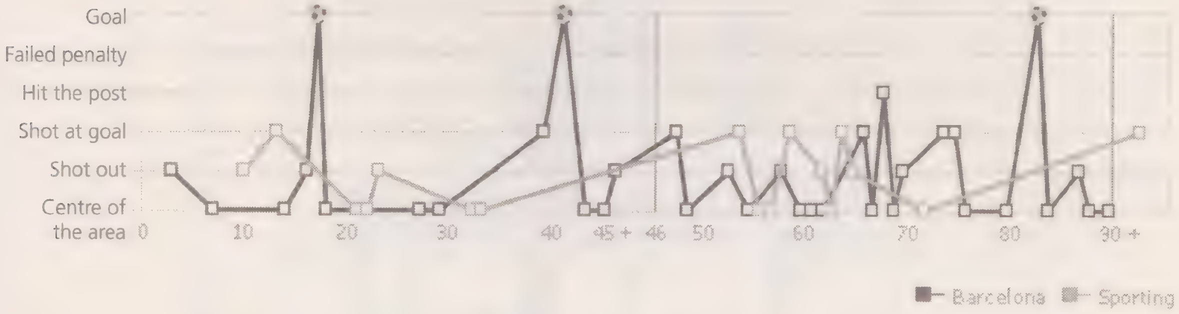
Beneath the chart, we have a stem and leaf diagram. In order to draw it, each value (numbers of beats per minute) is divided into two parts: the smallest numbers (in this case the units) will be the leaves, and the rest (tens and hundreds) form the stem. The lowest value is 48, then 54 and 54 again, then 3 58s, until the end which is a single 100. Note that the length of the list of corresponds with the ups and downs shown by the histogram. Therefore, the stem and leaf diagram shows what the histogram does, but it does more:

1. We can extract new information. By glancing at the histogram we can see that there is a value between 45 and 50, but we do not know its exact value, while

in the stem and leaf diagram that information is not obscured.

2. The stem and leaf diagram also allows us to see details that would otherwise go unnoticed. For example, it is unreasonable to think that all the students were taking the pulse for a minute. In that case, roughly half of them would be even and the other half, odd, but we can see that they are all even, which means that they counted for 30 or 15 seconds and multiplied by 2 or 4, and the result obtained in this way has a margin of error greater than if they had counted for a minute.

Sometimes specific graphs are designed for certain situations. An example of this are the graphs that appear in the online editions of some newspapers, accompanying their coverage of football matches. These graphs show the progress of the match, displaying each team's attempts on goal through a whole series of variables. In this way, they show everything from crosses into the box to goals scored and penalties conceded.

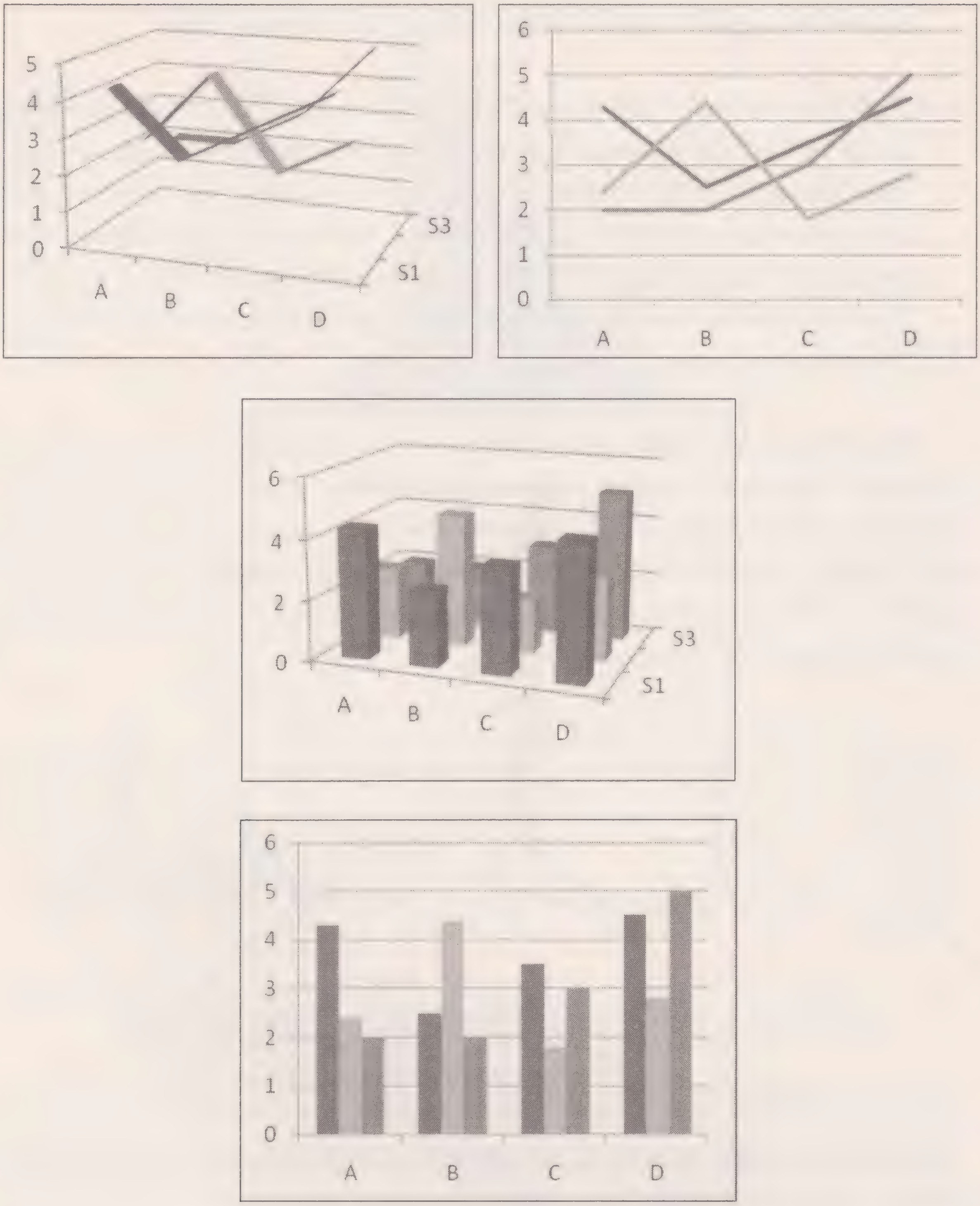


The attacking moves throughout a football match shown graphically.
(Source: Elpais.com).

However, it is normal to use a computer program for drawing graphs, generally with specific statistics software, spreadsheets or word processors.

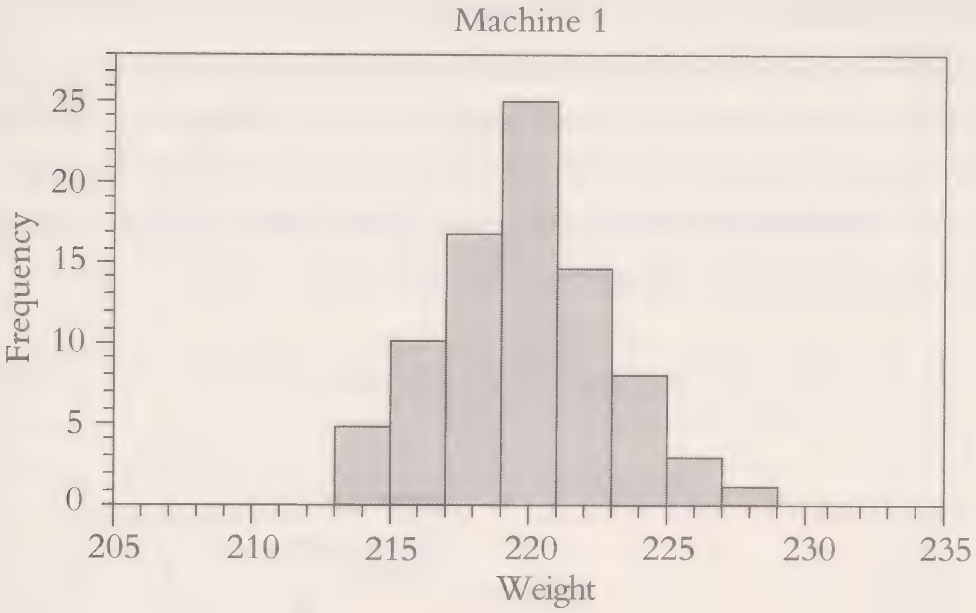
The word processing program that has been used to write these words allows the easy insertion of graphs. We can choose spectacular three-dimensional charts or simply 'flat' graphs, among other things. In general, three-dimensional graphs are the most eye-catching but least clear.

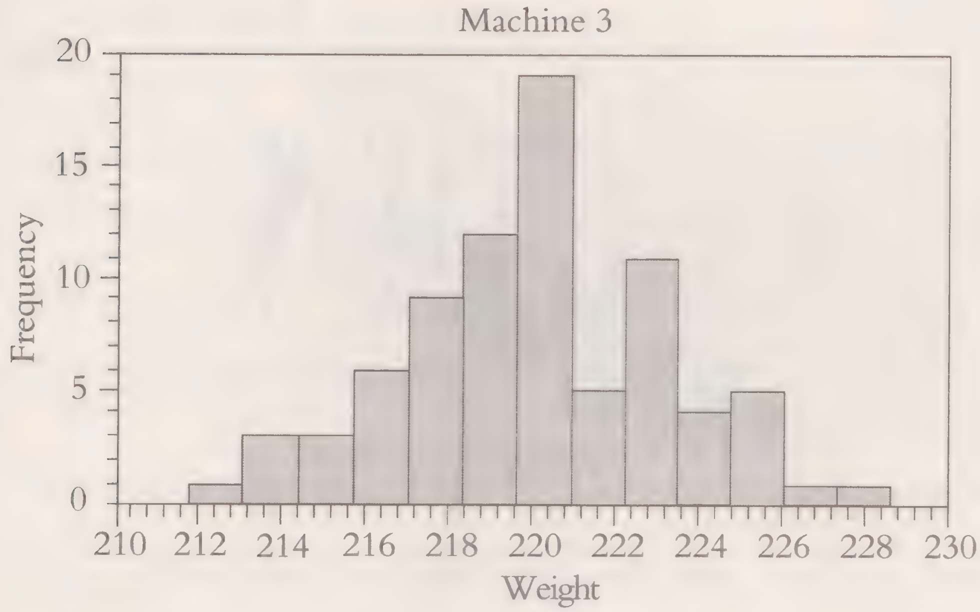
You can choose whichever your prefer, but you need to be sure that the information is properly conveyed to the reader.



Graphics created using Microsoft Word.

Let's go back to the example of the bakery in order to end this section on the graphical representations of variables. Suppose there is a third machine for which there is also a sample of weights for 80 loaves of bread (the same as the one for machine 1). How do you think the variability of this new machine compares to that of machine 1?





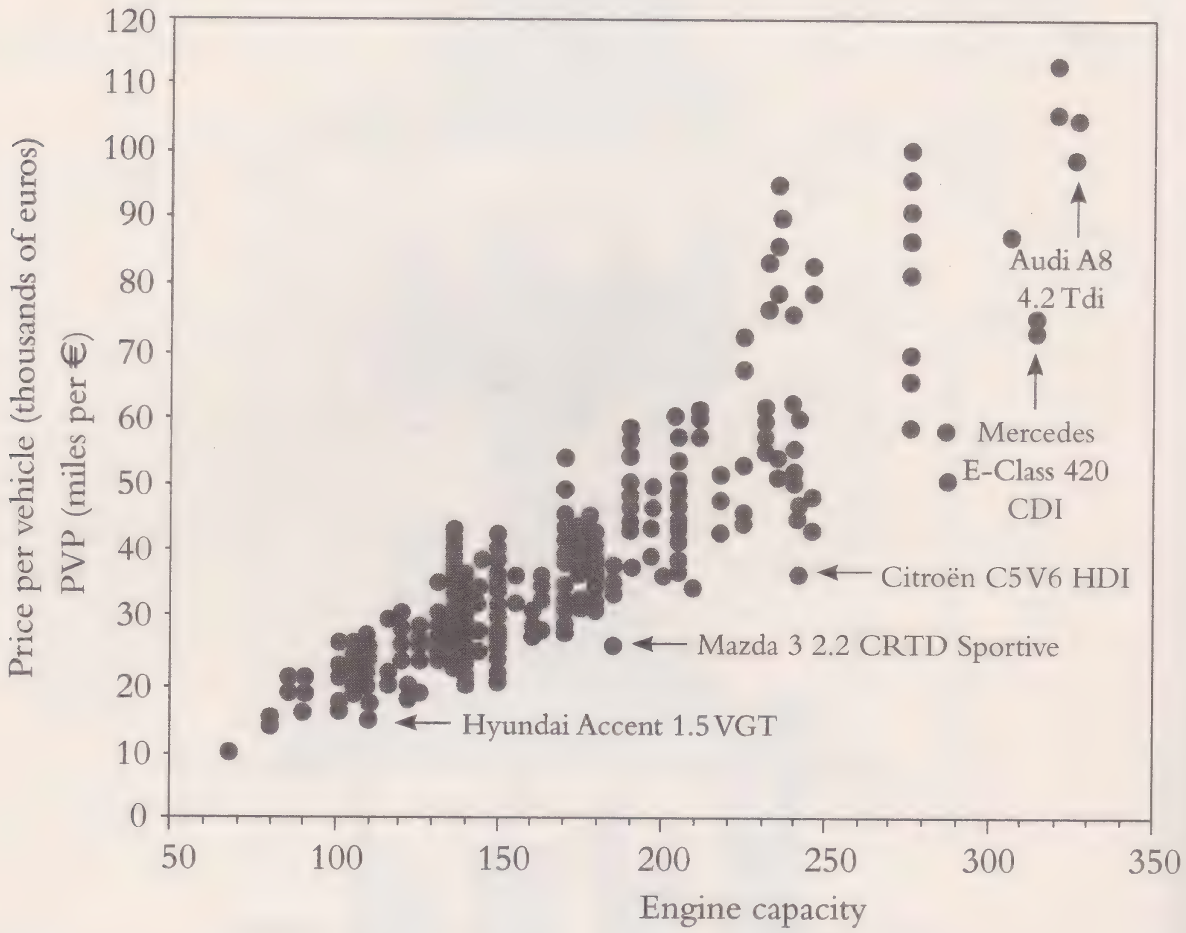
What would you say about machine 3 compared to machine 1?

If you thought that machine 3 produced more variability than number 1, you are wrong. The two histograms have been constructed with the same data set, but they look different because the scales are different.

You are right; that was a trick. The key lies in the way the data was presented. And the lesson learnt is that when you draw graphs to compare different situations make sure the scales are the same. The computer program adapts the size of the box to the variability of the data by default; you need to intervene and force the scales to be the same, otherwise it will lead to errors by people using the graphs, and you could even fall into your own trap.

Representations of relationships between two variables

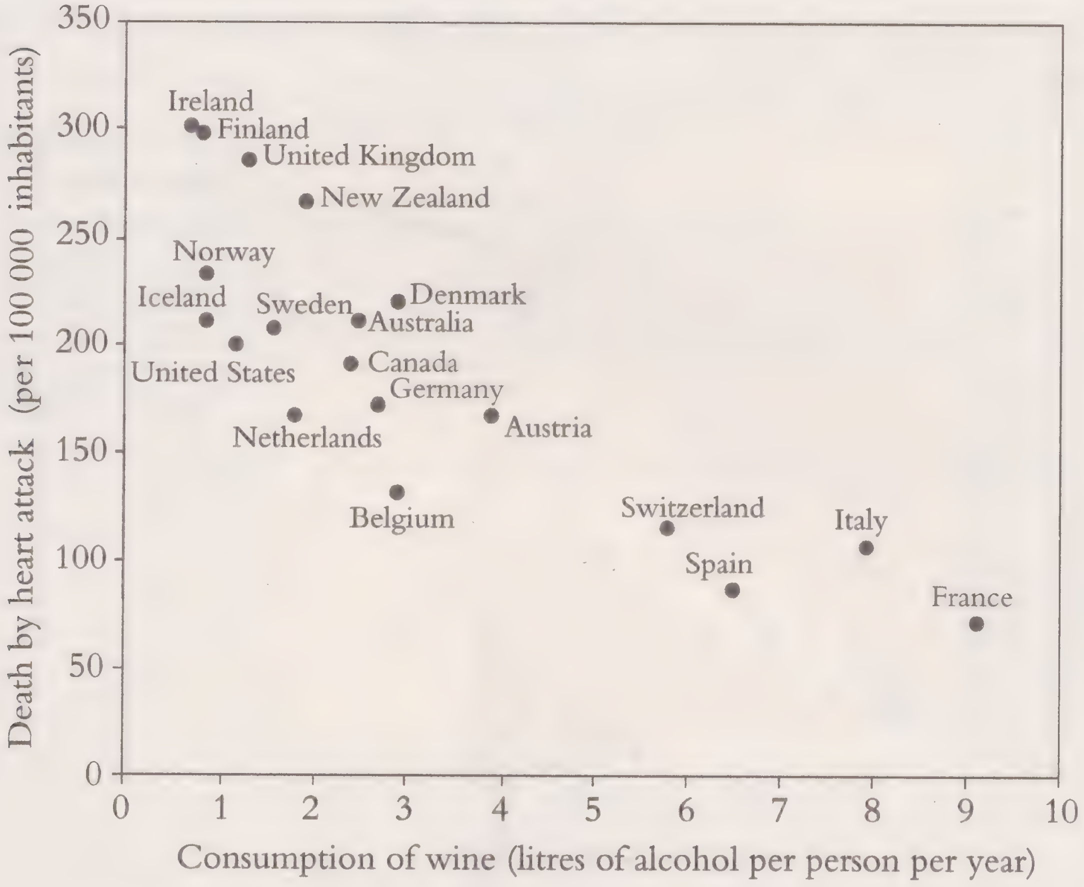
When we need to represent the relationship between two variables, graphs like the one below are used. This particular example shows the relationship between the price and power of a group of 449 diesel cars. You can see how some of the values of power, like 150 HP, are more frequent than others. You can also see that some cars are cheap compared to others of the same power.



The price–power relationship for a set of 449 diesel cars.
(Source: Real Automóvil Club de España web page, 10 November 2009).

Just because you see a close relationship between variables it does not mean that there is necessarily a cause-effect relationship between them. For example, if we create this type of graph relating fire damage to the number of firefighters that come to put the fire out, we would surely see that there is a close relationship between them: the more damage there is, the more firemen there are, but this does not mean the firemen caused the damage. In another example, primary school pupils with the largest feet make fewer spelling errors than those with small feet. Don't believe it? It's true – the oldest children have larger feet and make fewer spelling mistakes. In both cases there is a third variable that is maintaining the cause-effect relation between the two being analysed.

But there are cases where this is not so clear. On 28 December 1994, the *The New York Times* published an article on the health effects of drinking wine and included a table with the average wine consumption and the rate of death due to heart disease for a set of 21 countries. Below is a graphical representation of this data:

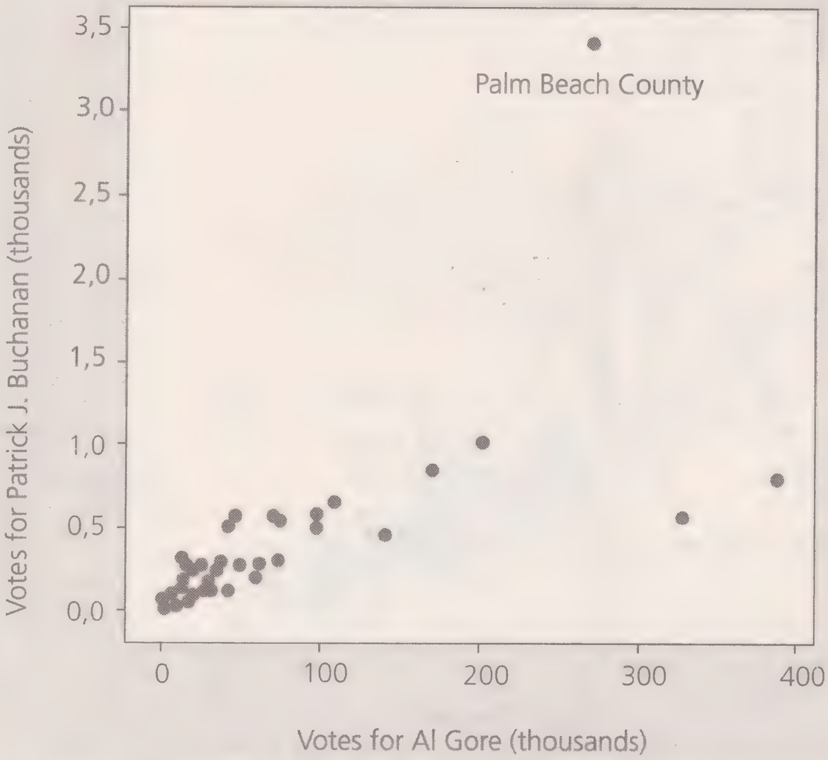


Ratio of deaths due to heart disease and consumption of wine in 21 countries. (Source: The New York Times, 28 December 1994).

We can see that the countries that consume most wine have a lower rate of deaths due to heart disease. But as we have said, this does not mean that there is necessarily a cause-effect relation between the two. This graph tells us that if we drink more wine (understood to be within reasonable limits) we have less risk of suffering from heart disease. The countries that consume the most wine are also the largest producers, and where wine is produced there is a climate, eating habits, customs... which could cause the decrease in this type of illness. It could also be due to the moderate consumption of wine, but this is not demonstrated in the available data.

SIMPLE GRAPHS FOR LEGALLY COMPLEX SITUATIONS

The presidential elections in the United States in the year 2000 with Democrat Al Gore against Republican George W. Bush had a very tight and also highly disputed result. In the State of Florida, with 6 million voters, Bush won by a margin of 537 votes, and whoever won that state had the majority needed for taking the presidency. There were objections to the count and the courts had to decide. Without considering the legal aspects, the graph shows that the votes obtained by Al Gore compared with the other candidate, Patrick J. Buchanan, in each of the 67 counties of the State of Florida.



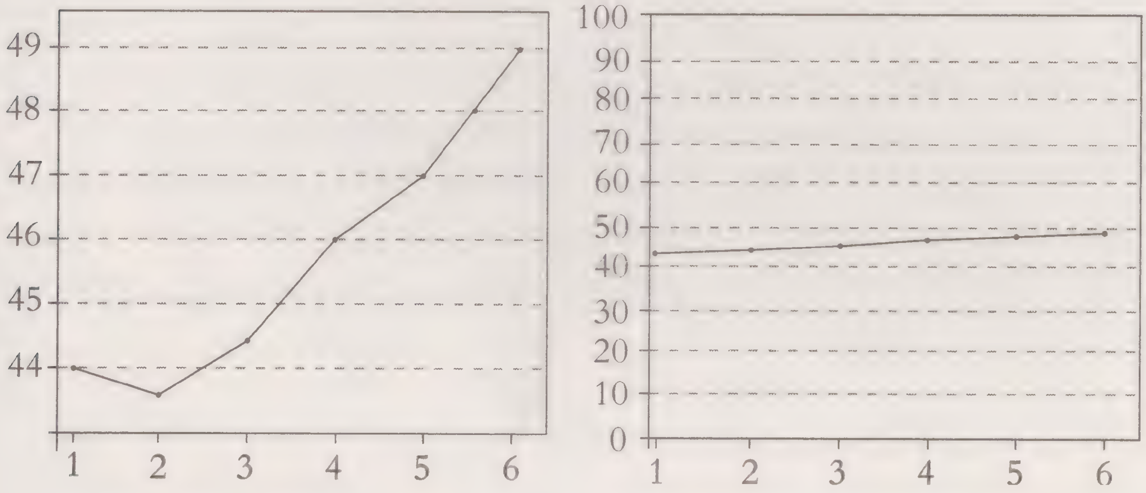
Patrick J. Buchanan's votes against Al Gore's in each of the 67 counties of the state of Florida (source: D.S. Moore: Learning from Data, in Statistics: A Guide to the Unknown, 4th edition).

The first thing that catches the eye is the value of Palm Beach, which does not follow the general pattern. The points are grouped together indicating a trend according to which Buchanan would receive around 1,500 votes in Palm Beach, but he got 3,411. Seeing this graph it is clear that something peculiar happened in Palm Beach. But there was no reason why Buchanan should receive a percentage of votes in that county far greater than in the others. He and his team declared that receiving one thousand votes was an optimistic expectation. It was soon clear that the peculiarity was the design of the ballot paper used to vote in that county. A hole had to be punched according to the chosen candidate, but the assignment of circles to each candidate caused confusion and many people (undoubtedly more than 2,000, as can be seen by looking at the graph) voted for Buchanan when they actually wanted to vote for Al Gore.

Warning: scales can catch you out

Given a set of data, the mean or standard deviation are concrete values. If someone tells us that the arithmetic mean of a data set is 3.1 and someone else insists that it is 4.2 one of them is wrong (or it could be that they are both wrong) because a set of data has one sole value for the mean. But this does not occur in graphical representations. Given a set of data, if we represent it on a histogram, the shape of the graph will depend on the scales chosen (we have already seen this with the data for hypothetical machine 3 in the bakery), and also with the widths used to express intervals and the limits of the intervals. Even if the width is the same, the histogram does not have the same shape when the limits are 190, 192, 194... as when they are 191, 193, 195.

For example, the development of an economic indicator for the last six months could be represented by the left-hand graph, which shows a spectacular rise, or the one on the right, in which it seems to have remained practically stable. The difference is the vertical scale.



The two graphs represent the same values, but the one on the left gives us the impression that there was a huge rise, while on the right-hand one it looks like the values have remained practically stable.

The horizontal scale can also produce surprises. The diagram on page 42 shows the annual sales of a product over the last four years, but as this graph was made in May 2010, the total data from that year only goes up to April. Although that is shown on the scale, the impression is that sales are falling, while (assuming that up to April a third of the annual sales have been made) the expectation is that sales for that year would be more than 150.

THE *CHALLENGER*

At some point we have all seen the image of the *Challenger* space shuttle in take off position: a kind of plane. standing vertically, attached to a large fuel tank which has something that looks like a small tank on both sides – the rockets that will put the machine into orbit. These rockets, just like other parts of the shuttle, cannot be transported in one piece, so they are manufactured in parts, transported to the launch site and assembled there. To ensure that there are no leaks in the joints, which could cause a catastrophe, large 6mm-thick rubber O-ring seals, 12m in diameter, are used.

On the night of 27–28 January 1986 a group of technicians and directors of the company which manufactured the rockets held a teleconference with their counterparts at NASA in order to discuss the possibility of postponing the launch scheduled for the next day. They were worried because the expectations for the ambient temperature at launch time were much lower than normal (between 26 and 29°F; −2 and −3°C) and they were worried that at these temperatures the joints would not guarantee a seal. They had data from previous launches, as they recovered the shells of the rockets and analysed them meticulously, and sometimes they had detected imperfections in the joints, although there had never been a serious accident. After analysing the available data it was considered that there was no evidence that the temperature would affect the possible deterioration of the joints and the decision was taken to go ahead with the launch.

The next morning, 59 seconds after beginning the launch sequence a flame began to come out of a joint which did not appear to be sealed. The flame grew rapidly until it reached the liquid fuel tank, which exploded, causing the death of the seven astronauts on board, and leading to uproar around the world and a general review of NASA's programme.

President Ronald Reagan ordered an investigation committee to be set up, made up of prestigious members of the scientific and space community. The commission determined that a very poor analysis of the available data had been made, and that one of the errors was not considering the flights in which the joints had not suffered any damage (figure 1) while a detailed analysis of the behaviour of the joints in all launches would have uncovered the relation between the observed imperfections and the launch temperature (figure 2).

Figure 2 clearly shows that there is no experience and, therefore, no assurance that there would be no problems at the expected temperature. Also, it can be seen that as the temperature decreases, more problems tend to appear. In figure 3 the number of joints with any damage (the magnitude of the deterioration is not defined) has been substituted by an evaluation carried out by the investigation committee; here the relation is even clearer. This is a clear example of how a simple graphical analysis of data can provide a lot of information on the problem that is being analysed.

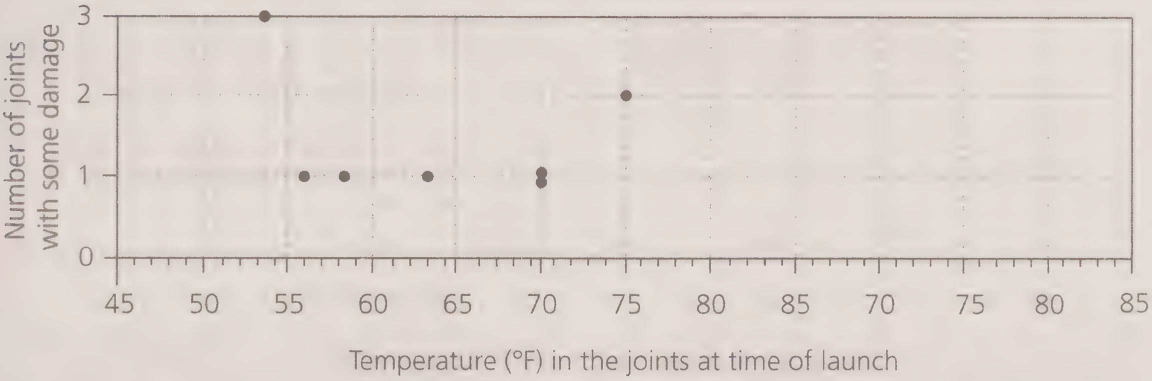


Figure 1. Each point represents a launch in which some damage to the joints had been detected.

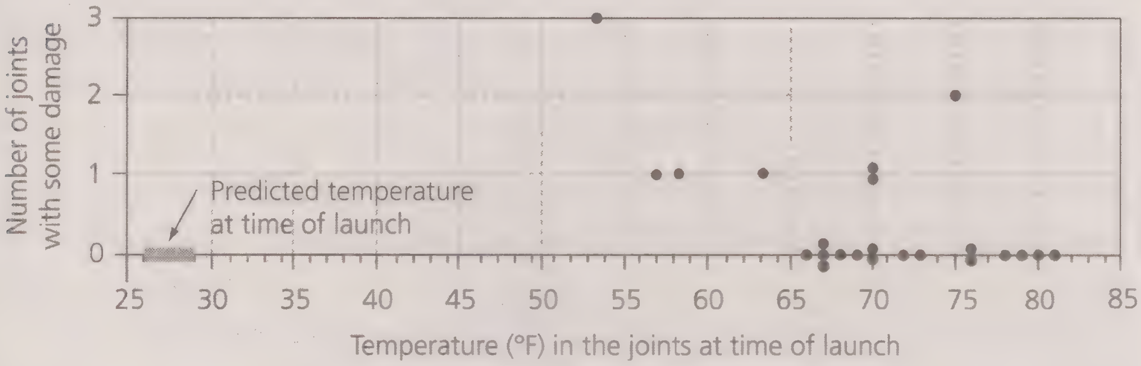


Figure 2. The scale is increased to incorporate the expected launch temperature; the flights with no imperfections in the joints are also included.

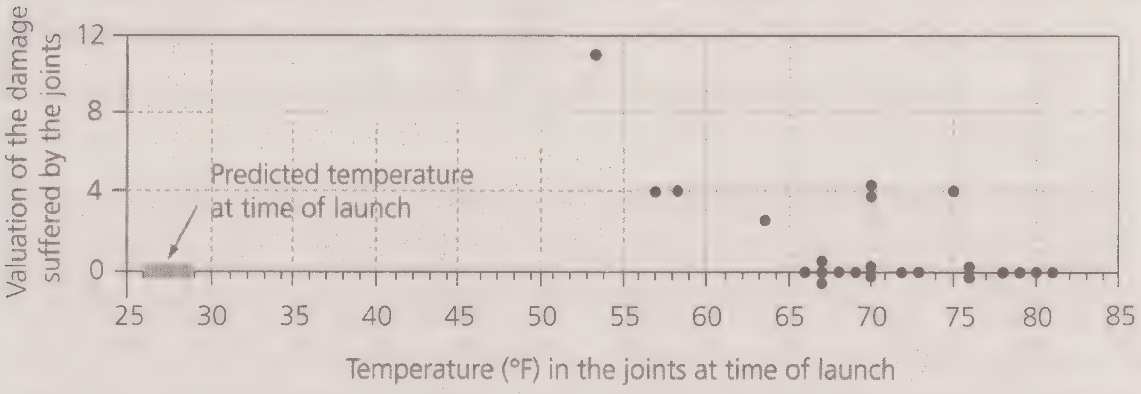
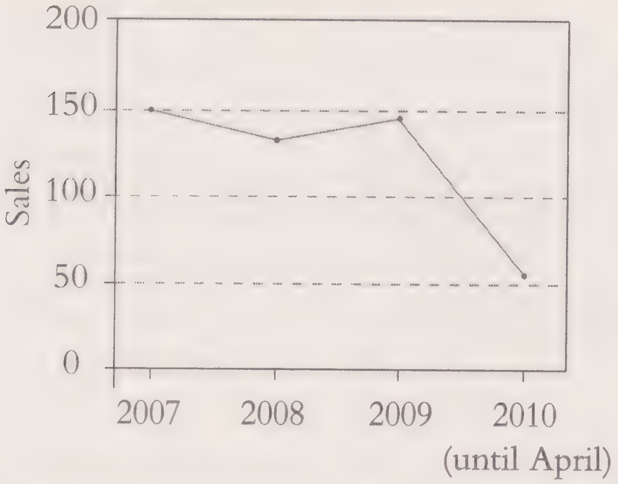
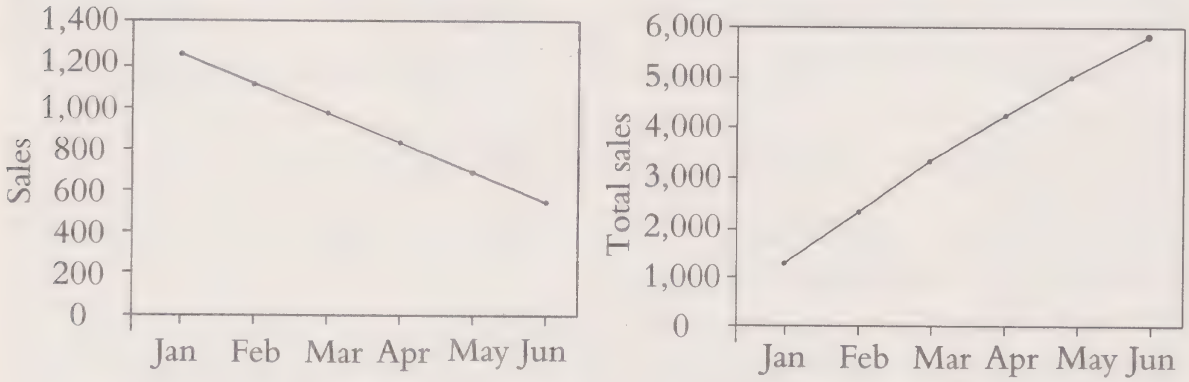


Figure 3. For each flight the damage suffered by the joints has been evaluated obtaining the evaluation which appears on the vertical axis (source: E.W.Tufte: Visual Explanations).



The four values are not comparable: the 2010 value only accounts for a quarter of a year.

Graphs can also give a different impression according to which variable we choose. For example, if your company is selling less and less, as shown in the left-hand graph below, you could produce the graph on the right representing the accumulated values which are still increasing.



Two ways of showing the development of sales: monthly (left) or accumulated (right).

But please do not go away with the idea that graphs are just shapeless forms that can be changed in order to show any idea we want. Clear and useful graphs which allow information to be understood at a glance can be designed, such as the histograms in the case of the bakery, and tangled, confusing and even unjust graphs can be created by playing with the scales or the variable they represent, or using confusing drawings and illustrations. In general, a little attention, critical assessment and a bit of experience help to uncover those situations.

Chapter 2

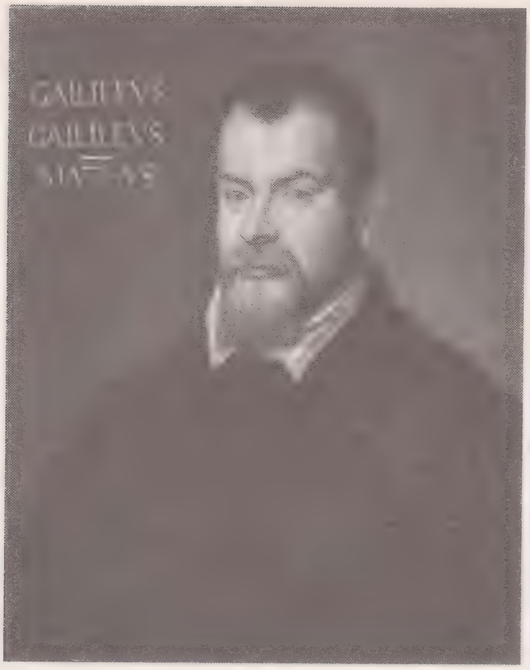
Calculation of Probabilities: Getting by in a World of Uncertainty

The calculation of probabilities aroused great interest among those who thought that the discipline would reveal strategies for winning in casinos, lotteries and other types of gambling. But it was soon discovered that it was of no use for that; actually, it aids those who design the games but not the player – as long as the former does his work properly.

Apart from gambling, calculating probabilities is useful in many fields. For example doctors use it to evaluate whether a massive vaccination programme would be effective. Industry uses probabilities to ascertain the quality of a large quantity of a commodity by inspecting just a few samples. The chances are the rest will be of the standard.

The calculation of probabilities from a mathematical standpoint came late in the day – already into the 17th century. Laplace's formalisation of probability as the number of favourable cases divided by the number of possible cases did not arrive until 1814, more than 2,000 years after Archimides discovered the formula for the volume of a sphere, which is far less intuitive. The prevailing idea prior to this was that results that depend on luck are unpredictable, have no rules and, therefore, are outside of man's ability to comprehend. It was also considered that chance was the work of the gods and was a magical property of divine design. Thus, its study was seen as dangerous territory for god-fearing mathematicians.

One of the works that is considered to be a pioneering study of the laws of chance was carried out by Galileo in around 1620 on behalf of an aristocrat. The object of the work was to find the most probable sum resulting from throwing three dice. It was thought that the values 10 and 11 were the most probable, but no one was quite certain and therefore they turned to one of the greatest scientific minds of all time.



Portrait of Galileo by Tintoretto. The Italian sage carried out one of the first studies in the field of calculating probabilities.

Galileo wrote a four-page report on his conclusions and how he had arrive at them. The reasoning behind it was the following:

1. A die has 6 faces and due to its symmetry we can assume that the probability of each one being thrown is the same. Therefore, the probability that one specific value comes up is 1 in 6.

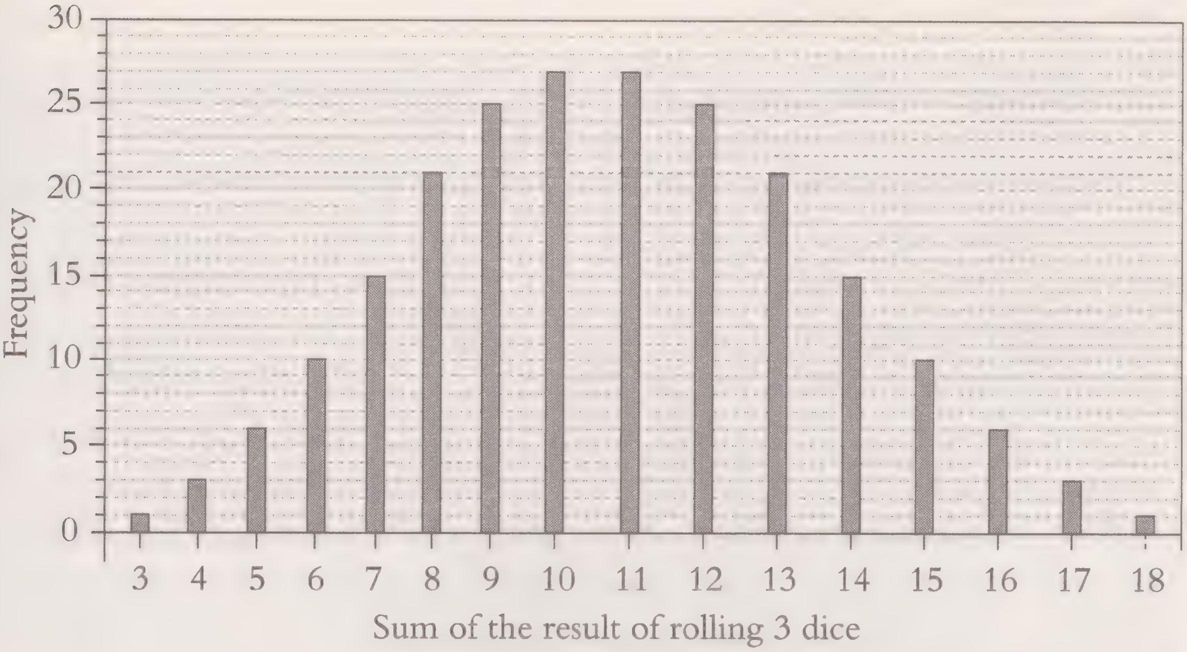
2. For each of the 6 possible results from the first die we have another 6 when we throw the second. In total, there are 36 possible results, as indicated in the table below, in which D1 is the throwing of the first die and D2, the second.

D1	D2	D1	D2	D1	D2	D1	D2	D1	D2	D1	D2
1	1	2	1	3	1	4	1	5	1	6	1
1	2	2	2	3	2	4	2	5	2	6	2
1	3	2	3	3	3	4	3	5	3	6	3
1	4	2	4	3	4	4	4	5	4	6	4
1	5	2	5	3	5	4	5	5	5	6	5
1	6	2	6	3	6	4	6	5	6	6	6

All doubles have the same probability of coming up, but not all the values of the sum appear with the same frequency. There is only one possibility from among 36 that they add up to 2 (rolling 1 and 1), and there is also only one that makes 12 (6 and 6), but there are 6 out of 36 (in other words, 1 in 6) that add up to 7, which is the most probable result.

3. If, instead of 2 dice, 3 are rolled, the reasoning is the same. For each of the 36 possible outcomes of rolling 2 dice there are 6 possibilities for the third, therefore the number of possible results is $6 \cdot 6 \cdot 6 = 216$. The fol-

lowing diagram shows the frequency with which each of the possible sums appears. The probability of a 10 or 11 being thrown is the same: 27/216 = 0.125, while the probability of a 9 or 12 coming up is slightly lower: 25/216 = 0.116.



The players must have had surprisingly fine perception to have noticed that the values 10 and 11 had the same probability, and that it is slightly higher than that for the values 9 and 12.

Calculation of probabilities and statistics

Statistics had been a discipline dedicated to collecting and describing demographic data of interest to the state's administration. Then in the 19th century, incorporating the calculation of probabilities opened a new, much broader array of possibilities. Soon insurance companies began to use mortality statistics and the theory of probability to estimate life expectancies and to tailor their insurance premiums accordingly.

Likewise, when electoral polls are taken, probabilities are used to turn the answers to the surveys into a prediction of the result of the election itself – as well as to evaluate how much confidence we should have in that prediction. In the same way, when evaluating the effectiveness of a new medicine by studying its effects on a sample of patients, conclusions are drawn using statistical methods that rely on how probable any side-effects will be.

But you do not need to be an expert or be capable of solving complicated probability calculation problems in order to understand and apply the most common statistical

GAMBLING AND THE ORIGIN OF CALCULATING PROBABILITIES

The calculation of probabilities is not just unusual because it took so long to be tackled by mathematicians, but also because of the events that led to the birth of that field. When studying advances in science we find selfless workers who sacrificed their lives in order to understand how the world works, or to improve the health and well-being of humankind. However, the calculation of probabilities arose from the interest of men of leisure who were interested in knowing the best strategies for winning at gambling, something which evidently took up a lot of their time.

One of the first discussions on the calculation of probabilities in mathematical terms can be found in the correspondence between Pierre Fermat and Blaise Pascal in 1654 regarding a problem set by a philosopher and gambler from the same period, known as the Chevalier de Méré. The problem was to find the fairest way to distribute the sum of a bet if the game had to be interrupted before the end; for example, if the bet is won when someone wins 3 games, but the game ends at 2 - 1.

Pierre de Fermat,
1601–65.

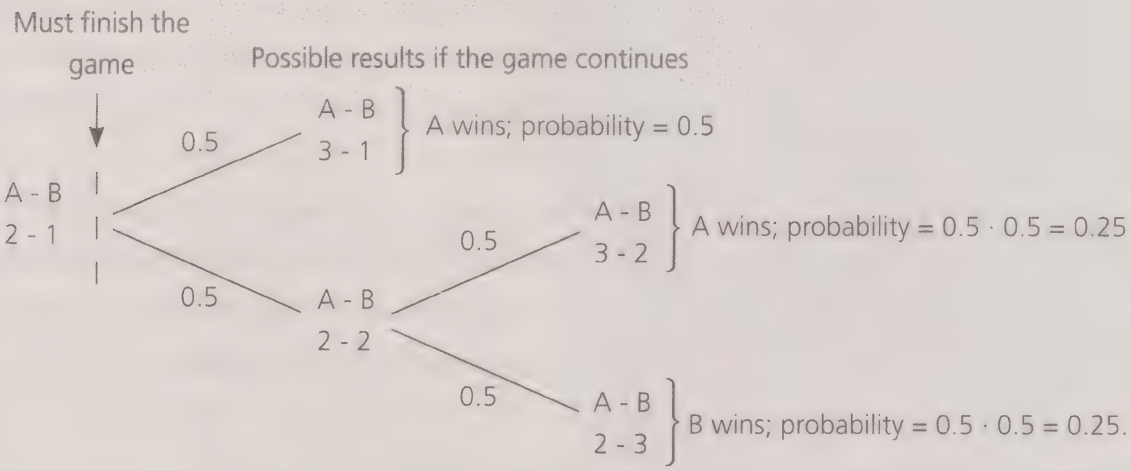
Blaise Pascal,
1623–62.

methods. Nor should we only relate statistics to casinos and gambling. Sometimes we can find roulette, dice and packs of cards on the covers of statistics books, but not forests, surgery, children, schools and production lines, where the application of statistics is much more useful and interesting.

One option would be that whoever is winning gets everything, another would be to share it out equally, but both Fermat and Pascal agreed that in a case such as this the most reasonable solution is that the player who has won 2 games takes three quarters.

The players are A and B and it is player A who has won 2 games, the reasoning is as follows: let's suppose that they continue to play and the probability of winning a game is 50%, the same for both players. The game would finish in one of the following ways:

1. The next game is won by A. As it would now be 3 - 1 the game would end, A wins and takes all the money. The probability of this occurring is 0.5.
2. The next game is won by B, leaving it at 2 - 2 and they continue playing. Next A wins, which makes it 3 - 2 to A and the game ends. The probability of this outcome is 0.5 x 0.5 = 0.25 (B wins and A wins).
3. The next game is won by B and then B wins again. In this case they finish 2 - 3 and B wins the game. The probability of this outcome is also 0.5 x 0.5 = 0.25.


Must finish the game
Possible results if the game continues
A - B
2 - 1
0.5
0.5
A - B
3 - 1
A wins; probability = 0.5
A - B
2 - 2
0.5
0.5
A - B
3 - 2
A wins; probability = 0.5 · 0.5 = 0.25
A - B
2 - 3
B wins; probability = 0.5 · 0.5 = 0.25.

In summary, if they continue to play, the probability of A winning would be 0.75 (0.5 + 0.25), while the probability of B winning would be 0.25. A would win 3 out of 4 times; therefore, it is reasonable for A to keep three quarters of the bet.

Probability and its laws

According to the ideas in Galileo's texts, if a test has n possible results, all of which are equally likely, and event A appears in k of the possible results, the probability of A occurring is:

$$P(A)=\frac{k}{n}.$$

In other words:

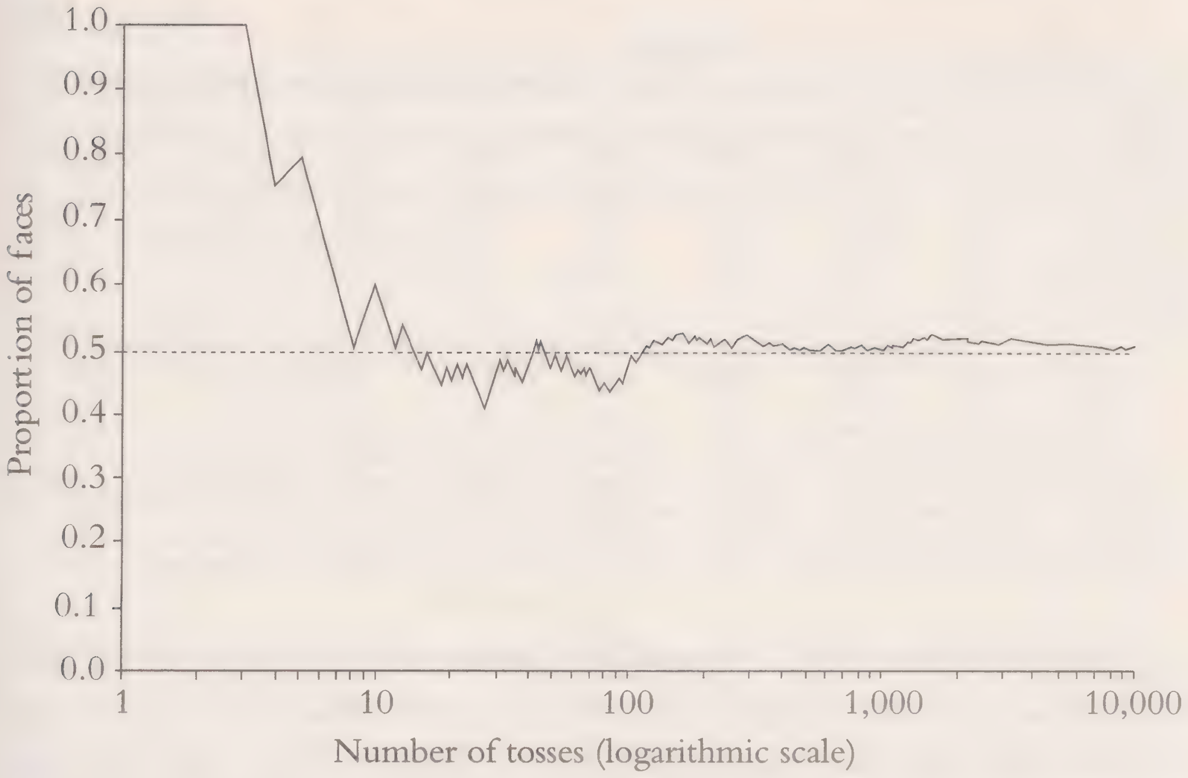
$$\text{Probability of an event occurring }=\frac{\text{Favourable outcomes}}{\text{Possible outcomes}}.$$

For example, if there are 5 balls in a bag, of which 3 are blue and 2 are black, if one is taken out at random the chance of it being blue is 3/5. It is that easy.

In some cases the theoretical probability can be defined by focussing on the symmetry of the object generating the results, as in the cases of rolling dice and tossing coins. Another approach consists of considering the probability as the proportion of times that the event occurs by indefinitely increasing the number of experiments. Thus, in order to know the probability of heads when tossing a coin it needs to be thrown many times while recording the number of heads. The same occurs with dice; when we say the probability of a certain value is 1/6 we are referring to a perfect dice, and perhaps that is not what we have in our hands. There is only one way to find out.

Some researchers have tossed coins and rolled dice many times, noting the results as they did so. One of them was English mathematician John Kerrich, who was imprisoned in Denmark during World War II. As he was in prison he tossed a coin 10,000 times: he got 5,067 heads and 4,933 tails.

The proportion of heads fluctuated as indicated in the figure below, although these results are not Kerrich's but are a simulation. As the number of throws increases the fluctuations are dampened and it is reasonable to assume that the proportion will have a constant value if the throws continue indefinitely. That value will be the probability of getting tails with that coin.



The proportion of heads when tossing a coin 10,000 times (obtained by a simulation).

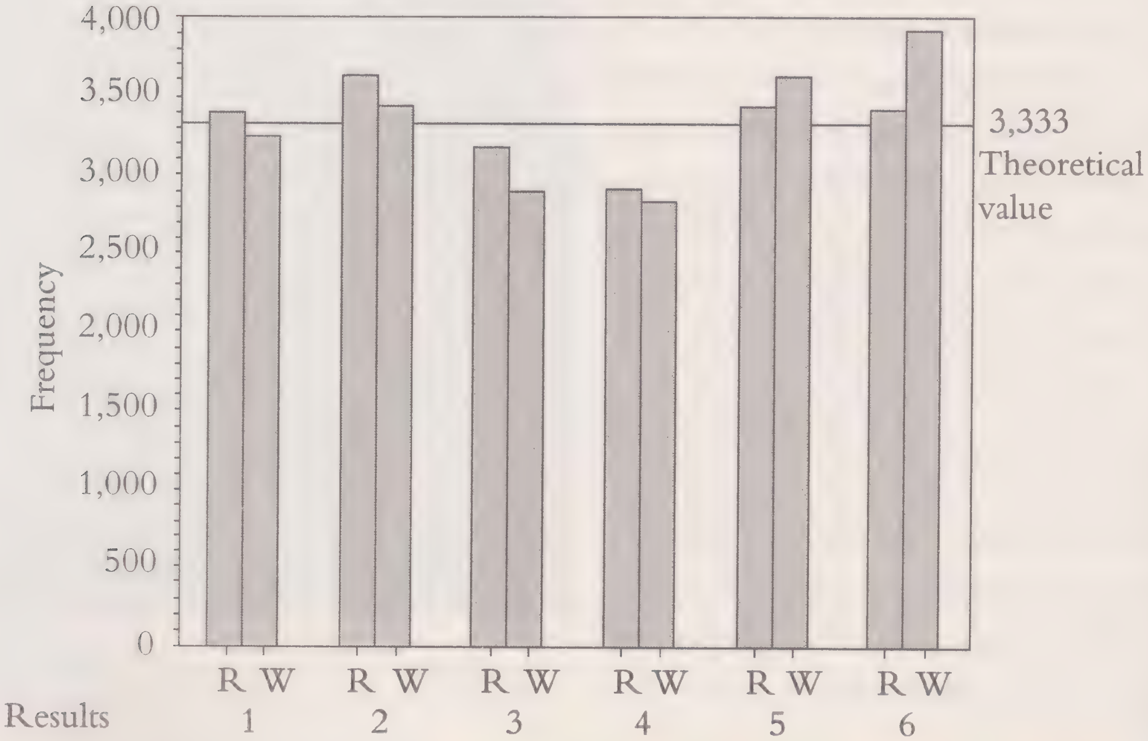
Other researchers who have carried out similar exercises are the Comte de Buffon, an 18th-century French scientist who obtained 2,048 heads throwing 4,000 times, and Karl Pearson, one of the fathers of modern statistics tossed a coin 24,000 times (him or one of his assistants) and obtained 12,012 heads.

In dice throwing, the most famous results were obtained by a Swiss astronomer called Wolf when he threw two dice no fewer than 20,000 times, one red one and one white one. The results obtained from the test are listed in the table on the next page.

The Comte de Buffon made several probability studies in the 18th century. This portrait is by François-Hubert Drouais.

		White die						Total	Proportion
		1	2	3	4	5	6		
Red die	1	547	587	500	462	621	690	3,407	0.170
	2	609	655	497	535	651	684	3,631	0.182
	3	514	540	468	438	587	629	3,176	0.159
	4	462	507	414	413	509	611	2,916	0.146
	5	551	562	499	506	658	672	3,448	0.172
	6	563	598	519	487	609	646	3,422	0.171
Total		3,246	3,449	2,897	2,841	3,635	3,932	20,000	1.000
Proportion		0.162	0.172	0.145	0.142	0.182	0.197	1.000	

The results obtained with the coins are consistent with the supposition that they are well balanced (the probability of heads = 0.5), but in results from throwing dice the probabilities are quite far removed from their theoretical values. Both the white and the red dice seem to have a deficit for the values 3 and 4. Let's take a look at the results on a graph in order to see them more clearly (R = red die, W = white die). In Chapter 3 we talk about testing hypotheses by statistical means and discuss whether it is reasonable to consider that the dice are not balanced.



Results obtained by throwing one red die (R) and one white die (W) 20,000 times.

The 'or' rule

The probability that event A 'or' another event B happens, if they cannot both occur at the same time, is the sum of each of their probabilities. For example, in a pack of 52 poker cards (without jokers) the probability of choosing an ace or a picture card when selecting one at random is:

Probability of an ace: $P(A)=\frac{4}{52}$ ((favourable cases/possible cases).

Probability of a picture card: $P(B)=\frac{12}{52}$.

Probability of an ace or a picture card: $P(A \text{ o } B)=P(A)+P(B)=\frac{4}{52}+\frac{12}{52}=\frac{16}{52}$.

The 'and' rule

The probability that event A 'and' another event B happen, if they are independent, in other words, if the occurrence of one does not affect the probability of the other, is equal to the product of their probabilities. For example, when throwing a die twice, the probability of getting a 3 and then a 4 is:

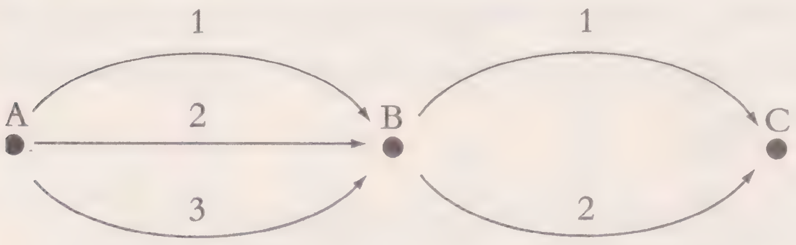
Probability of a 3 $P(A)=\frac{1}{6}$ (favourable cases/possible cases).

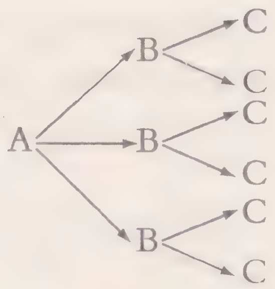
Probability of a 4 $P(B)=\frac{1}{6}$.

Probability of a 3 and then a 4: $P(A \text{ y } B)=\frac{1}{6}\cdot\frac{1}{6}=\frac{1}{36}$.

Counting cases

Counting the favourable cases or the possible cases can be the most laborious part of the work, although in some situations the counting can be done by simple reasoning or by referring to similar situations. For example, if to get from A to C you have to pass through B and there are 3 routes to get from A to B and 2 to get from B to C, how many ways can we get from A to C. For each of the three options that exist for getting to B we have 2 for getting from B to C, then in total there are 6 different ways of getting from A to C.





Let's have a look at another case which seems more complicated. In the football pools there are 3 options for each match: home win (1), draw (X) or away win (2). What is the probability of winning the pools with 14 matches.

It is clear that there is only one favourable case: there is only one winning combination. The possible cases seem more difficult to count, but we can use the same idea as the one for counting the routes to get from A to C: the first match has three possible results, for each result of the first there are three options for the second; in other words, if there were only 2 matches the possibilities would be $3 \cdot 3 = 3^2$. Following the reasoning we reach the conclusion that for 14 matches the possible cases are 3^{14}. So, the probability of getting 14 right if the pools are filled in at random is $1/3^{14}$, approximately 1 in 4.8 million.

Combinatoric formulae are also very useful in these cases; we will look at some in the context of the problems provided later.

Applying the rules

Below, the rules set out above are applied to an example. We are going to calculate the probability of getting 3 heads and 2 tails, in any order, when tossing a coin 5 times. As you will soon see, this problem is more relevant than it may appear at a glance. Let's break it down:

1. The probability of getting a head, and also that of getting a tail, is 0.5.

2. The probability of getting 2 heads in 2 throws is $0.5 \cdot 0.5 = 0.25$. We have applied the 'and' rule because the results are independent, in other words, if a head comes up first it does not increase or decrease the probability of a head coming up second.

FRANCIS GALTON AND THE QUINCUNX

Francis Galton (1822-1911) was a scientist with a broad spectrum of interests, from anthropology to economics, philosophy, meteorology and statistics. He was Charles Darwin's cousin. He had a comfortable private income which allowed him to dedicate his time to the things he was interested in, travel among them. Although he studied medicine he barely worked in the profession and as soon as he received his family's inheritance he went off to explore the world. Among other adventures he spent two years exploring Africa and was granted a gold medal from the Royal Geographical Society in recognition of his activities.

Among his contributions are a detailed analysis of fingerprinting and his recommendation that it be used as a system for identifying offenders, as has been the case ever since. He also studied inheritance and the transmission of characteristics, observing that children of tall parents also tend to be tall, but not as tall as their parents, and the same occurred with the children of short parents, they tend to be short but less so than their parents, producing an effect of returning to the population average which Galton called 'regression' toward the mean, coining a new term which is one of the key words in current statistics.

In order to illustrate the way in which the variability presented itself due to random causes, he invented a device called 'quincunx' in which balls pass through an area of nails fixed to the quincunx in such a way that the balls hit them and randomly fall to the right or to the left. After this junction, the balls fall into channels and it can be seen that the arrangement formed by them complies with a Gaussian bell. The quincunx is still used in today's class roooms to illustrate normal distribution, and nowadays simulators can also be found online.

3. The probability of getting: head and head and head and tail and tail in five throws: $0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 = 0.5^3 \cdot 0.5^2 = 0.03125$. (It is true that we could have also written 0.5^5, but in order to make it applicable to other cases that we will see below it is in our interest to separate the probabilities of heads and tails here too).

What we have calculated is the probability of getting 3 heads (H) first and then 2 tails (T) in this order: H H H T T. We want to calculate the probability of getting 3 heads and 2 tails in any order, in other words, getting H H H T T or T T H H H, or H T H T H etc.

Order	Probability
H H H T T	0.5^3 x 0.5^2
H H T H T	0.5^3 x 0.5^2
H T H H T	0.5^3 x 0.5^2
"	"
"	"
T T H H H	0.5^3 x 0.5^2

The probability we are looking for will be the sum of the probabilities of each of the possible orders. We can add by applying the 'or' rule because the events are incompatible (two different orders cannot occur at the same time). And as the probabilities are the same for all the orders we can also multiply the probability of having a specific order by the number of orders that may occur. (We are now entering the field of combinatorics). If we have *n* objects, they can be sorted in *n!* different ways. For example, if we have 5 books and 5 spaces to place them in the library, the first could be placed in any of the five spaces, for the second we can only choose from 4, 3 for the third, 2 for the fourth and there is only one left for the fifth; so the possibilities are: $5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$. In our case we also have 5 'objects' but they are not all different. We have 3 the same on one side and 2 on the other, so the permutations between these equalities need not be taken into account so we divide by 3! and 2!. The number of combinations that can be made from 3 heads and 2 tails is:

$$\frac{5!}{3! \cdot 2!} = 10.$$

Now we have everything we need to calculate the probability we are looking for. It is:

$$\frac{5!}{3! \cdot 2!} 0.5^3 \, 0.5^2 = 0.3125.$$

And what use is it to know the probability of getting 3 heads when throwing a coin in the air 5 times? It is clearly of little use, but below you will see that the process we used can be generalised for other fields which are hopefully of interest.

Chance has families

On 29 April 2004, a reader sent the following question to a computer help column in a widely circulated newspaper: "I have used an Excel spreadsheet to calculate random numbers using the function '=rand()', but it always gives me small, near zero, results and I want a system for doing the lottery, which involves 6 numbers below 50."

It seems like this reader thought that if a value depends on chance (if it is random) it is not bound by any kind of rules and it is already written, but this is not the case. There are different types of random variables, and the first dichotomy we can draw is between continuous variables (they are measured on a continuous scale) such as weight, length, density, etc. and discrete ones (with values that are separate from one another) such as the number of defective pieces in a batch, the number of cars arriving at a petrol station per minute. Actually, we have a whole catalogue of distributions of probability and the first thing we do when faced with a random variable is see if it fits into any of the distributions described. Normally it does fit, so that there is no need to deduce the formulae for calculating the probabilities, or its mean or other characteristics. Someone has already done it and the information is available to us to apply.

At first it can seem difficult to distinguish the different types of random variables, like someone who does not know a particular genre of music and they find it difficult to identify the types of songs, although with a little practice it is easy to distinguish one from the rest. Below we will analyse some characteristics and uses of the three probability distributions, undoubtedly the best known examples. First we'll look at two discrete ones and then, a continuous one.

Something we have already seen: binomial distribution

Applying the general rules of the calculation of probabilities we have used the following expression to determine the probability of getting 3 heads and 2 tails (in any order) when throwing a coin 5 times:

$$\frac{5!}{3!\cdot 2!} 0.5^3 \cdot 0.5^2.$$

In general, the number of successes when carrying out *n* experiments when the probability of success, *p*, is constant it is a random variable that follows a very well-known probability distribution known as 'binomial'. There is no need to deduce new formulae to calculate probabilities when we have a case that fits this scenario.

A VERY USEFUL FORMULA

Beyond specific numbers, if we try to calculate the probability of getting *x* heads with *n* throws and calling the probability of getting a head *p* and 1 – *p* for tails, the formula is:

$$\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}.$$

The interesting thing is that this formula is not only valid for the problem of tossing coins, but it can also be generalised for any field which fits the following table:

Tossing coins	Generalisation
Tossing *n* coins.	Carrying out *n* experiments.
Heads or tails may come up.	Each experiment has two possible results, which we will call 'success' and 'failure'.
Both the probability of heads (*p*) and tails(1 –*p*) is constant for all throws.	Both the probability of success (*p*) and failure(1 –*p*) is constant in all experiments.
The probability of getting *x* heads in n throws is of interest.	The probability of getting *x* successes in *n* experiments is of interest.

Let's have a look at these three problems:

1. A production line produces 1% defective pieces. If the pieces are placed in 50 unit boxes, what is the probability of a box having 2 defective pieces?

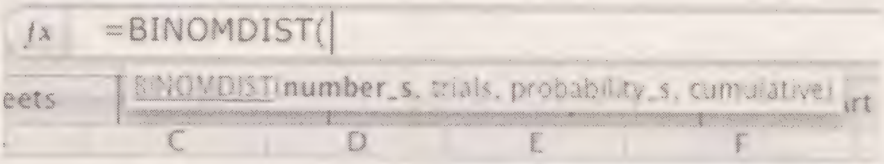
2. A professional basketball player has a success rate of 75% on free throws. What is the probability of netting 8 if he throws 10?

3. If a couple has 4 children, what is the probability of them having 2 boys and 2 girls?

What do these problems have in common? All three fit the scenario described and, therefore, they are very easy to resolve.

Problem	Number of experiments	Probability of success	Number of successes *x*	Answers
Defective piece	50	0.01	2	$\frac{50!}{2!\cdot 48!}0.01^2 \cdot 0.99^{48} = 0.076.$
Basketball	10	0.75	8	$\frac{10!}{8!\cdot 2!}0.75^8 \cdot 0.25^2 = 0.282.$
Children	4	0.5	2	$\frac{4!}{2!\cdot 2!}0.5^2 \cdot 0.5^2 = 0.375.$

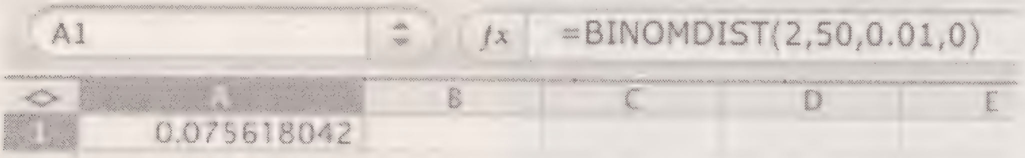
The calculations can be done with a spreadsheet. In Excel it is written exactly as shown below:



=BINOMDIST(

BINOMDIST(number_s, trials, probability_s, cumulative)

The last value, after the probability of success, indicates whether we only want to calculate the probability for the number of successes indicated (for example, having exactly 2 defects in the first scenario) and in this case we put a 0, or the accumulated probability up to that value (2 defects or less) in which case we put a 1.



=BINOMDIST(2,50,0.01,0)

	A	B	C	D	E
1	0.075618042				

In the case of the basketball player we have to suppose that the probability of scoring is constant, in other words, it does not depend on pressure from the crowd, nerves or the state of the game (surely one of the virtues that a good player should have is insensitivity to these things). In the case of the sons and daughters many people think that if they have 4 children it is more likely that 2 of them are boys and 2, girls, but the probability of this occurring is only 38%; the most likely is any other combination.

From death by horse-kick in the Prussian army to goals in the football league: Poisson distribution

When a variable fits the binomial model both the number of occurrences and non-occurrences can be counted (the number of correct pieces and defective pieces) and there is also a maximum number of occurrences. For example, the maximum number of correct pieces will be the total of those we have.

Sometimes we are met with variables that represent the number of events that occur per unit of time, or in space, such that its non-occurrence cannot be counted and there is no limit, at least from a theoretical point of view. Typical examples of this type of variable are: the number of daily visits to a website, the number of breakdowns suffered by a lift in a year, the number of calls taken by a telephone switchboard at lunchtime and the number of emails you receive daily. For occurrences in space, the examples could be the number of rust points per metre of steel cable, the number of defects per square metre (or for every 10 square metres) of cloth, or the number of raisins in a spoonful of breakfast cereal.

In 1837 French mathematician Siméon Poisson searched for a way of modifying the binomial distribution formula to adapt it to situations such as these and found a surprising expression for which you only need to know the average number of occurrences (λ) in order to calculate the probability of any number of them happening. The formula for the probability of x occurrences is:

$$P(x) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

19th-century French mathematician, Siméon Poisson.

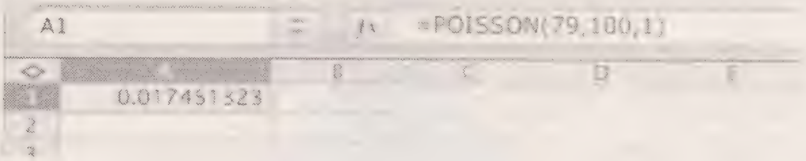
So, if a lift breaks down on average twice a year ($\lambda = 2$), the probability that it does not break down during the year is:

$$P(x = 0) = e^{-2}\frac{2^0}{0!} = 0.14.$$

And if a web page receives an average of 100 hits per day (let's assume it is the same every day of the week, although really we should have to distinguish between working days and days off), the probability of it receiving less than 80 hits in one day is:

$$P(x < 80) = \sum_{i=o}^{79} e^{-100}\frac{100^i}{i!}.$$

Which is not exactly comfortable to calculate, but that is what spreadsheets are for:



```
A1        fx  =POISSON(79,100,1)
0.017451523
```

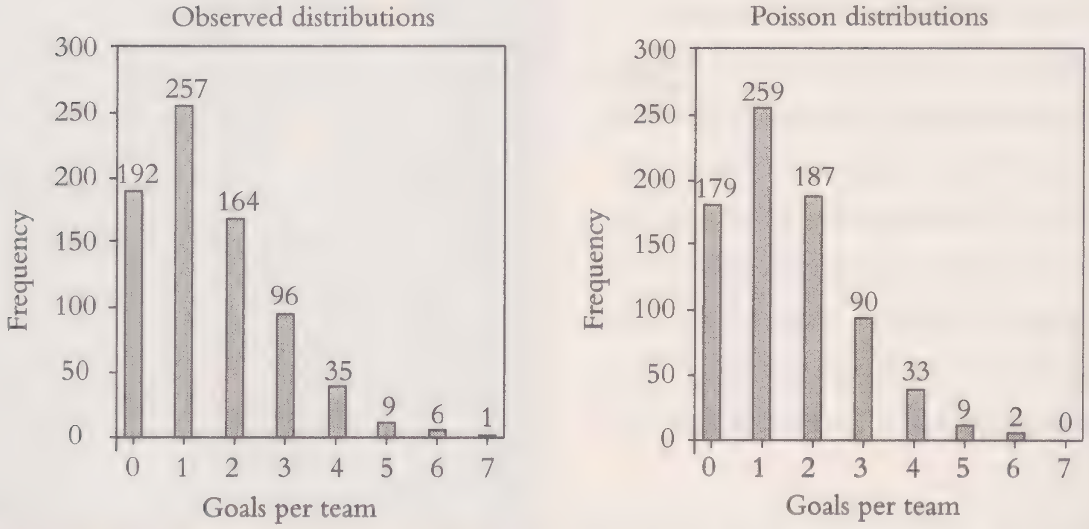
In 1898 Russian economist and statistician Ladislaus Bortkiewicz published a book showing that Poisson distribution could be used to explain the statistical regularity which can be seen in the occurrence of unusual events. He used data on suicides and accidental deaths in various circumstances, but his most famous example is that of the number of soldiers killed by horse-kicks in the 14 Prussian army corps over a period of 20 years (from 1875 to 1894).

Ladislaus Bortkiewicz, a Russian statistician who bettered the use of Poisson distribution.

In the table below the frequency corresponds to the number of army corps multiplied by the number of years in which the number of deaths indicated occurred (total = 14 · 20). The average number of deaths per army corps and year is (91 + 2 · 32 + 3 · 11 + 4 · 2)/280 and using this value in our formula we get the theoretical frequencies indicated in the following table.

Number of deaths by horse-kick	Observed frequency	Theoretical frequency
0	144	139
1	91	97
2	32	34
3	11	8
4	2	1
5 or more	0	0
Total	280	280

Given the choice of searching for data more applicable to our time, we could consider the number of goals a team scores in a football match, as this variable fits Poisson's distribution plan well. There are events per unit time (per match), there is no limit and the number of 'not goals' cannot be counted. The diagram on the left shows the number of goals scored in each of the 380 matches of the Spanish football league in the 2008-2009 season. The diagram on the right shows the data deduced from our formula.

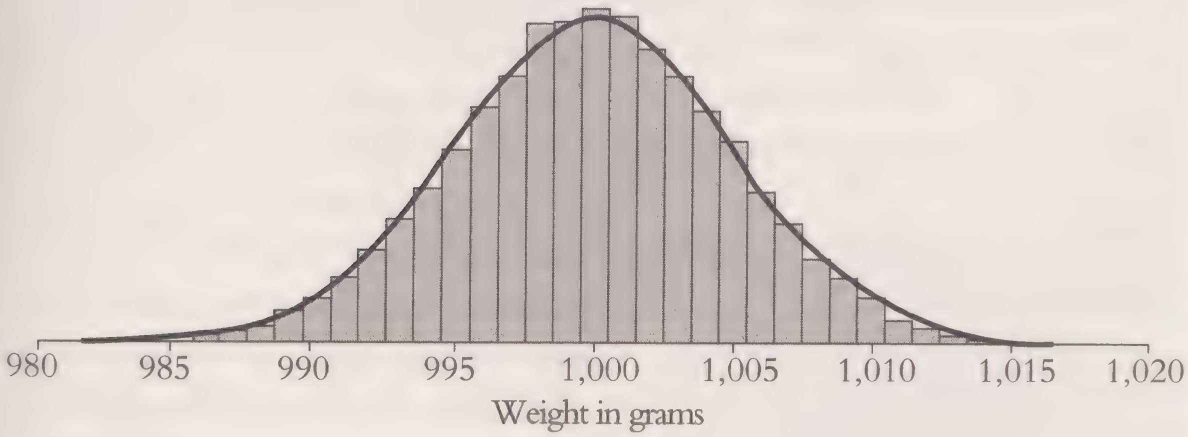


Observed distribution and theoretical distribution according to Poisson's model for the number of goals scored by each team in the 380 matches of the 2008-2009 season of the Spanish football league.

Indeed, the profiles are very similar. The Poisson model explains the variability of the number of goals scored by a team in a match.

The Gaussian bell or normal distribution

The Gaussian bell is very popular in mathematics. Its shape corresponds to the profile of the histogram that represents a large set of values affected by what is called natural variability. For example, 1kg packs of sugar do not weigh exactly 1,000g, some weigh a little more and others a little less. This is an inevitable variability, produced by numerous small causes, undoubtedly imperceptible on their own, but together they can have a notable effect. The graph below shows that most of the values are grouped around the central value and as we move away from that value the observations become more and more scarce. This is the typical shape of the Gaussian bell or normal distribution.



Possible distribution of the weight of a large number of 1kg packs of sugar, which describes a typical Gaussian bell.

The mathematical expression for the shape of the bell was deduced for the first time in 1733 by French mathematician Abraham de Moivre, but its name is associated with the German Carl Friedrich Gauss, who in 1809 used it to explain his theory on measuring errors, especially those in astronomical observations. Gauss proved that regardless of whether the measured object is close or far away, or of whether it is large or small, if the measurements are repeated in the same conditions, the values obtained are always distributed in this peculiar way.

But normal distribution occupies a very important place in the theory of statistics not only because it explains the theory of errors, but also because it represents a type of variability that is common in nature – and in the human body in particular.

Image of Gauss on a German 10-mark note. In the centre a graph of normal distribution can be seen.

One of the names linked to the origins of modern statistics is that of Belgian scientist Adolphe Quételet who, during the 19th century, carried out numerous studies to demonstrate statistical regularity (number of crimes, number of births, number of deaths, etc.). In his search for data to demonstrate normal distribution, he found himself with an unexpected gift. A Scottish magazine had published the height and thoracic girth of more than 5,000 soldiers belonging to various Scottish regiments, and that data allowed him to demonstrate that the variability shown by the soldiers was the same type as that described by the law of errors.

Adolphe Quételet, one of the most important statisticians in the 19th century.

In the words of Quételet: "If a person little practised in measuring the human body were repeatedly to measure one typical soldier, 5,738 measurements made on one individual would certainly not group themselves with more regularity... than the single measurements made on the 5,738 Scottish soldiers. If the two series were given to us without being particularly designated, we should be much

STIGLER'S LAW OF EPONYMY

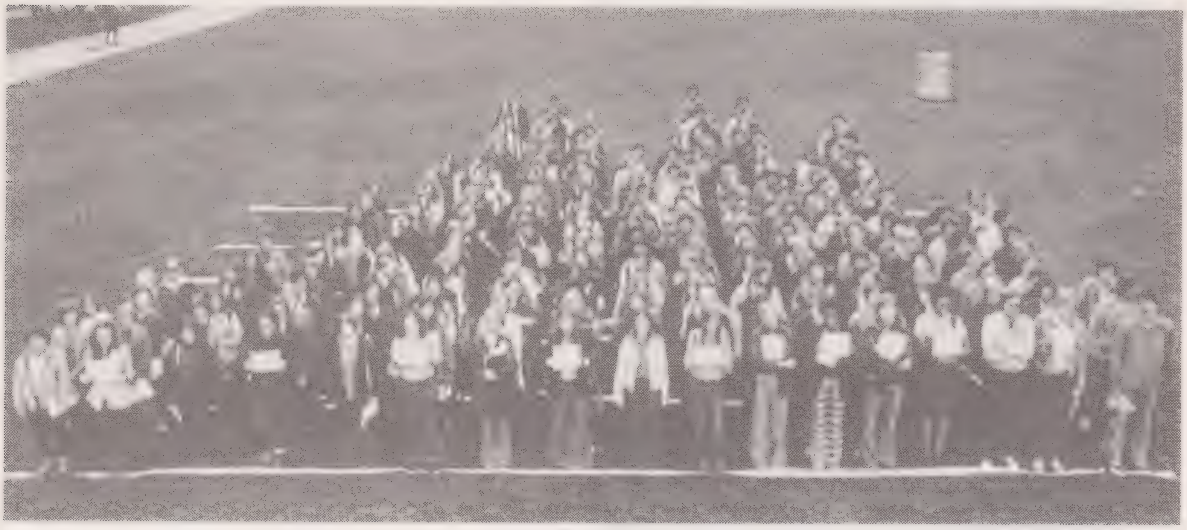
Portrait of Abraham de Moivre, who deduced the so-called Gaussian bell, many years before the famous German mathematician did.

Many scientific discoveries, diseases, laws, theories and constants, carry the name of the person who discovered them – for example: Alzheimer's, the Euler constant, Fermat's last theorem, Halley's comet or the Gaussian bell. The name given to this law or phenomenon is an 'eponym'.

Stephen Stigler, professor of statistics at the University of Chicago and recognised historian on the subject, has come up with a law which concisely states that "no scientific discovery is named after its original discoverer". Of those mentioned above, Alzheimer's disease (Alois Alzheimer) appears to have been described by at least half a dozen scientists before him; Euler's constant was discovered by Jacob Bernoulli; Fermat's last theorem, if it is a theorem, is not Fermat's (it would be a conjecture), as it was not demonstrated until 1995 by Andrew Wiles; Halley's Comet was discovered by astronomers before the birth of Christ, although it is true that Edmond Halley calculated its orbit and predicted the date of its return. Closer to our field of interest, it is well documented that normal distribution, with its bell-shape, was not discovered or first described by Gauss but by French mathematician Abraham de Moivre, who published his work on the matter in 1733, nearly 80 years before Gauss.

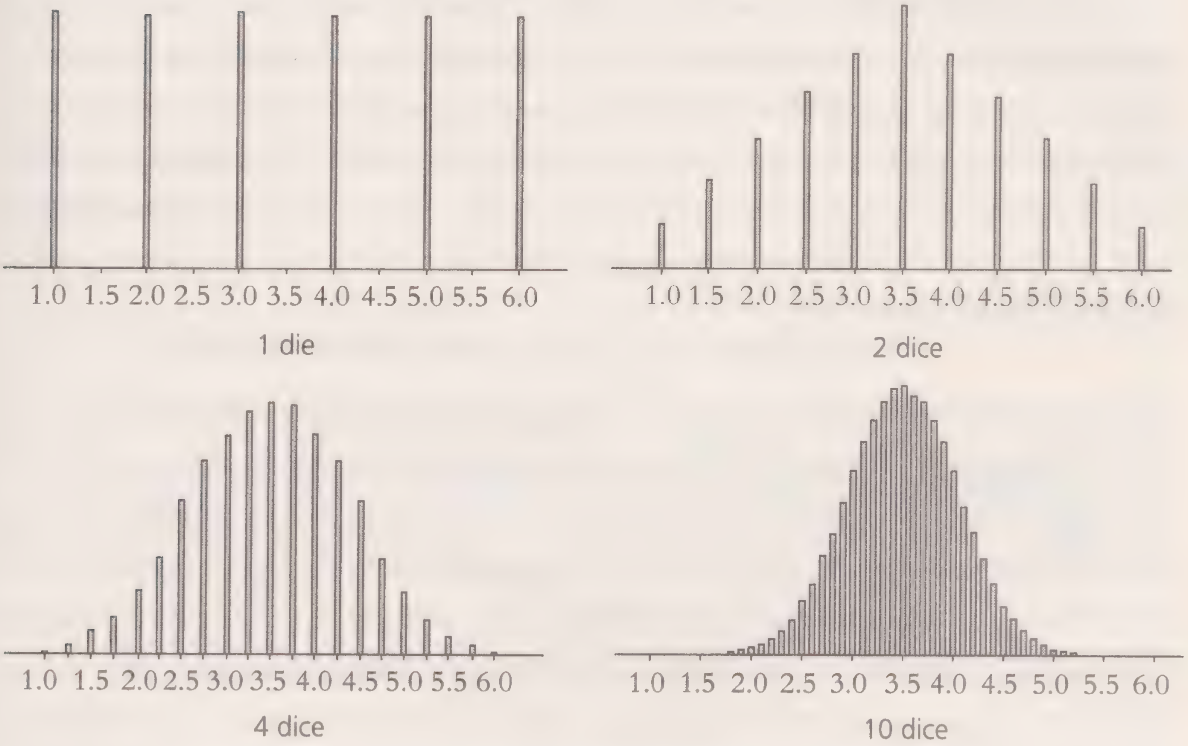
This does not mean that some scientists unduly take credit for the merits of others. What happens is that some make relevant contributions or uncover a subject that already existed but was not well-known, and from there, through no fault of their own, the discovery is associated with them. Professor Stigler published an article on this subject, which had already been mentioned by many others before him, among them Robert Merton, who is often quoted. But with a touch of humour, to add a new example to his law, he proposed that it should be called Stigler's law, and that is how it remained.

embarrassed to state which series was taken from 5,738 different soldiers, and which was obtained from one individual."



A living histogram; each individual is standing in the column corresponding to their height (source: Edward R. Tufte: The Visual Display of Quantitative Information, citing the work of Brian L. Joiner "Living Histograms", published in 1975 in the International Statistical Review).

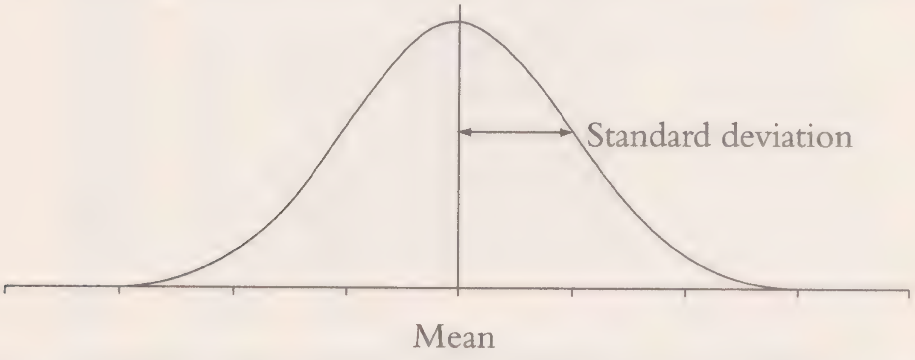
There is another reason for normal distribution's special importance. Often the prime material of the studies are the means: the *mean* production per plant using one fertiliser or another is compared, or whether the *mean* value of a sample correlates with the supposed value for the mean of the population is analysed, etc. So, the measurements show variability because, depending on the sample taken, the mean will be one value or another, and that variability can be characterised, for practical purposes, through normal distribution, even if the original data from which it is calculated is not normal. For example, when throwing a die we get values with a distribution that is nothing like normal distribution. It is a discrete distribution with only 6 possible values: 1, 2, 3, 4, 5 and 6, all of them with the same probability. If we throw two dice and take the mean, now not all values (of the mean) have the same probability; it is more likely that the mean is 3.5. If we do it with 4 dice the profile of the bar diagram that represents the probabilities of the values which the mean may take on it now reminds us of the shape of a Gaussian bell. If we throw 10 dice, which would be to take a sample size of 10, the profile of the bell is now evident. Thus, if we work with averages, our distribution is always normal.



The distribution of the means tends towards normal distribution even when the values which constitute the means are not normal.

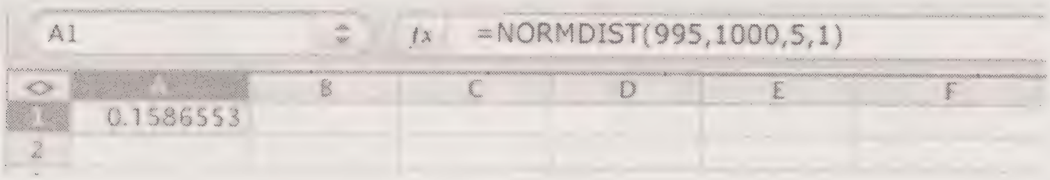
In any case, and despite its undoubted importance, the name 'normal' is a little unfortunate. When a distribution is called 'normal' it sounds as if the others are strange, but this is not the case. But the name has stuck and it is the most commonly used, although they are some who choose to call it 'Gaussian distribution'.

If, due to the nature of the data being worked with, it is considered that its variability can be characterised by means of that distribution (it can also be verified graphically or by means of the suitable statistical tests) it is sufficient to know two values in order to completely define it. The mean, which is the value at the centre of the bell, and its standard deviation, which indicates its flatness or slimness.



Mean and standard deviation: the two parameters that characterise normal distribution.

If the weights of the packs of sugar follow a normal distribution with a mean of 1,000 grams and a standard deviation of 5 grams, we can calculate what proportion of packs will have a weight over 1,010 grams, between 995 and 1,010 or less than 995. Until recently a few calculations had to be made and a few tables consulted (which are still included at the end of many books on statistics), but nowadays it is sufficient to use a spreadsheet. For example, the probability that a pack weighs less than 995 grams would be:

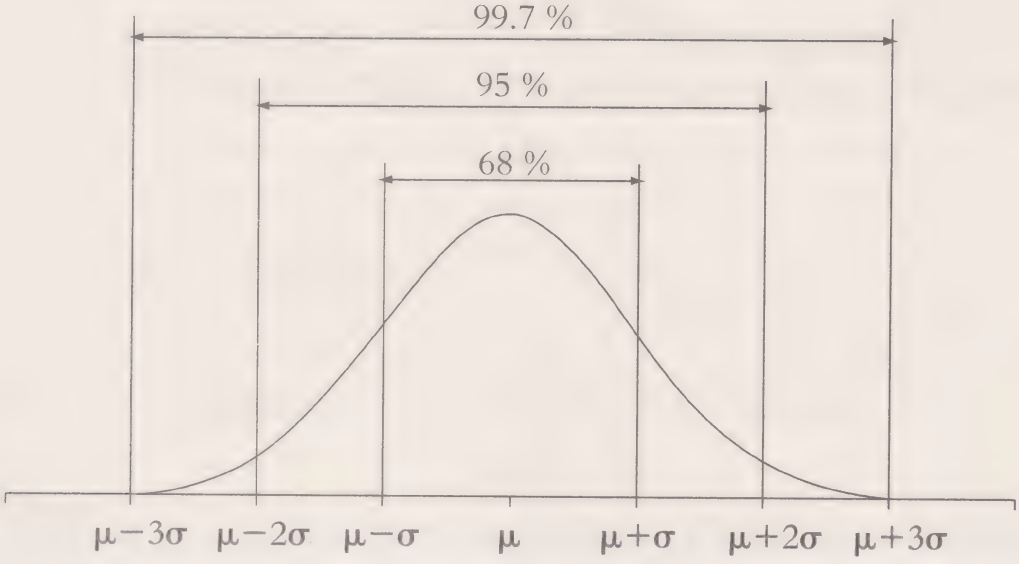


A1 fx =NORMDIST(995,1000,5,1)

	A	B	C	D	E	F
1	0.1586553					
2						

Note that it can be confirmed that approximately 16% of the packs will contain less than 995 grams, but nothing can be said about the weight of a specific pack. For the same reason we can talk of the life expectancy of a population, but not the age at which each individual will die.

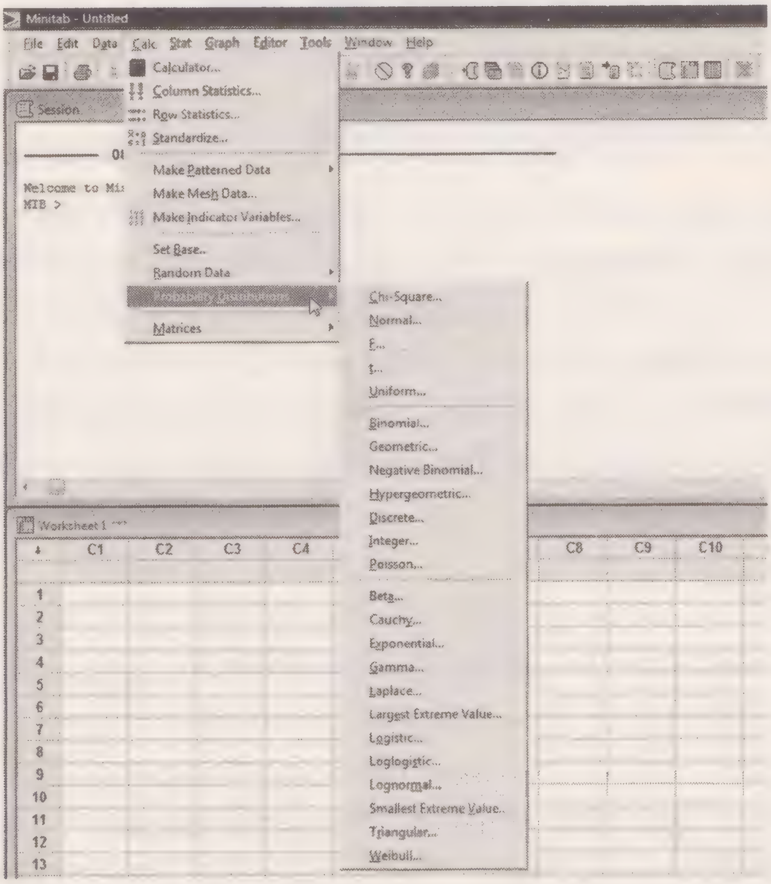
There are also a few quick rules that are based on the property that, regardless of the values of its mean (μ or 'mu') and its standard deviation (σ or 'sigma'), 68% of the elements are found in the interval $\mu \pm 1\sigma$, 95%, in the interval $\mu \pm 2\sigma$ and 99.7%, in the interval $\mu \pm 3\sigma$. So, in the above case, as the mean was $\mu = 1{,}000$ and the standard deviation, $\sigma = 5$, in the interval 995-1,005 we will find 68% of the observations; therefore, outside we will have 32%, 16% on each side, and so 16% will be below 995.



The rule is also useful for interpreting the value of standard deviation. If you think of the distribution of heights, the mean could be 170cm and the standard deviation must be between 6 and 7cm, as, looking at the high values, only 1 or 2% of the population is taller than 1.90m, which would be three standard deviations above the mean.

Other distributions: reflection on the 'theoretical' models

There are other probability distributions. For example, if the variable is continuous and all the values are equally probable, the distribution is called 'uniform'. Excel gives this distribution values between 0 and 1 when asked for random values with the instruction '=rand()', and these were those found by the person who wanted a number to fill in lottery tickets. The distributions catalogue includes much more; in the following diagram we can see those contained in statistics software package Minitab.



Probability distributions that can be calculated directly with the Minitab application.

Catalogued or not, the model should not be confused with reality. Although the sphere is a very common geometric figure in the universe, there are certainly no perfectly spherical objects. Then what is the point of formulae for the surface or volume of a sphere? They are of use because for practical purposes they give sufficiently close values. The same thing occurs for probability distributions.

One of the most useful examples for illustrating normal distribution is that of the heights of people, but if we had the exact heights of the millions of adults who inhabit the planet, we could prove that they do not exactly fit our well-known bell, and nor would they if we categorised them by sex, race, or any other characteristic.

It is a good reference model that allows, with all necessary accuracy, estimations to be made on values of heights, but it is still a theoretical model that does not exactly coincide with reality. The same thing happens with the other distributions in which, simply because the hypotheses considered do not exactly comply in practice, they are still theoretical models (calling a model theoretical is an unnecessary specification), but, they are enormously useful.

A bit of fun: surprising probabilities

Probability calculation problems are owed respect because they can be quite difficult, even though their wording may seem simple. (For example, what is the probability of the same winning combination of lottery numbers appearing on two occasions?)

The interesting thing is that surprising probabilities are found, in some cases very different to what we would intuitively expect. And we can use a little imagination when trying to solve difficult problems. Let us consider some examples.

False positives

In a routine check-up it is found that a person is suffering from an illness that affects 1% of the population. The type of analysis carried out produces 5% false positives – it indicates that they have the illness when in reality they are not suffering from it. What is the probability that they actually have the illness?

Perhaps you are thinking that it is 95%, but then you would be wrong, as in reality it is far lower. For every 1,000 people analysed there are 50 false positives (5%) and 1 true positive. Then, if among the 51 positives there is only one true one, the probability that it is theirs is only 1/51, in other words, slightly below 2%.

The birthday problem

In a class of 30 students, what is the probability that 2 or more have their birthday on the same day?

Most people would expect this probability to be very small, but the truth is that it is not that small at all. We start by calculating the probability that two people were not born on the same day. The first person has no restrictions, they could have been born on any day of the year (365 favourable cases over 365 possible cases), but the second has to have been born on any day other than the day on which the first person was born (364 favourable cases over 365 possible cases):

$$\frac{365}{365} \cdot \frac{364}{365} = 0.9973.$$

Similarly, the probability that 3 were born on different days would be:

$$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} = 0.9918.$$

And the probability that 30 were born on different days:

$$\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \ \ldots \ \cdot \frac{336}{365} = 0.2937.$$

There are only two possible outcomes: everyone is born on different days or at least 2 were born on the same day. Then the probability that at least two were born on the same day would be:

$$1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \ \ldots \ \cdot \frac{336}{365} = 0.7063.$$

That is to say, the probability that, in a group of 30 people, 2 or more were born on the same day is around 70%. With 23 people it is slightly over 50% and with 40 it is 89%.

SHARING BIRTHDAYS

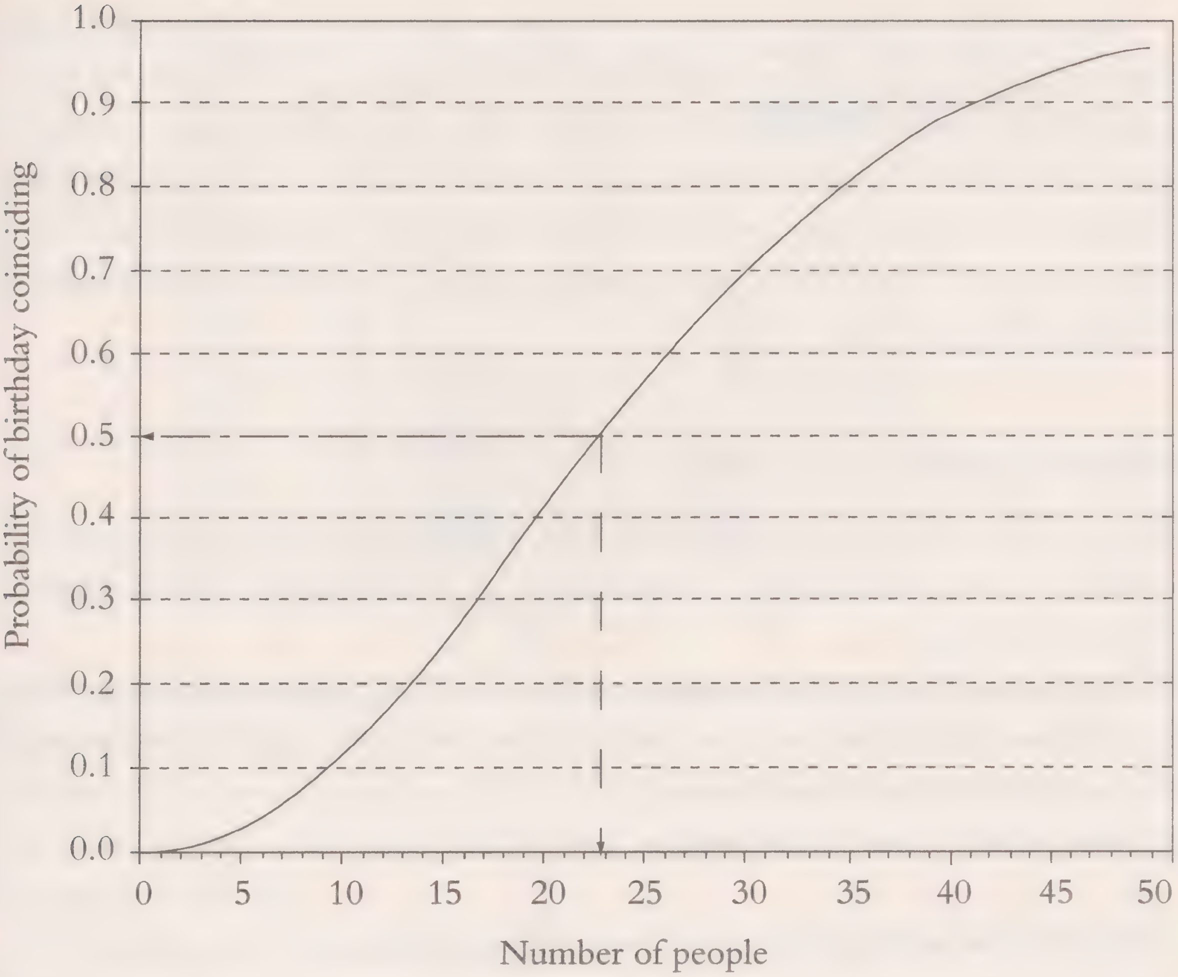
Although it may seem surprising, the probability that in a group of 23 people two or more have the same birthday is slightly greater than 50% (exactly 50.7%), as you can see from the graph, opposite. If you're not convinced by the reasoning behind calculating that probability, you can see what happens in various groups of 23 people. The problem is finding them and finding out their dates of birth. However, there are alternatives.

On a football pitch there are 23 people (11+11+1, the referee) and it is easy to find both the line ups for matches and the dates of birth of the players. Taking the matches from the first weekend of the 2012–13 season of the English Premier League, of the 10 matches played, in seven there were people on the pitch who shared birthdays, specifically:

Arsenal v Sunderland	No concurrences.
Everton v Manchester United	Pienaar (Everton) and Kagawa (Man Utd): 17 March.
Fulham v Norwich City	No concurrences.
Manchester City v Southampton	Nasri (Man City) and Puncheon (Southampton): 26 June.
Newcastle United v Tottenham Hotspur	Simpson (Newcastle) and Kaboul (Spurs): 4 January.
Queens Park Rangers v Swansea City	Cissé (QPR) and Graham (Swansea): 12 August.
Reading v Stoke City	Pearce and McAnuff (both Reading): 9 November.
West Bromwich Albion v Liverpool	Reid and Olsson (both WBA): 10 March.
West Ham United v Aston Villa	No concurrences.
Wigan Athletic v Chelsea	Luiz and Mikel (both Chelsea): 22 April.

So in this case there was a 70% 'success' rate, which is unusually high. Although the probability is 50% that there will certainly be at least one 'success', the probability of no matches having people with the same birthdays up to 10 matches with concurrences are as follows:

Matches with concurrence	0	1	2	3	4	5	6	7	8	9	10
Probability	0.001	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0.001



The probability that in a group two people were born on the same day of the year as a function of the size of the group.

Variant (the same but the other way round): in a group of 30 people, what is the probability that 2 or more die on the same day (although not necessarily in the same year)?

The winning combination comes up twice

Let's have a look at another case with surprising probabilities. A person has been playing the lottery for their entire adult life (let's say 50 years); if they buy 2 tickets per week, what is the probability that during that period the same winning combination comes up more than once.

Even though there are other variants, normally 6 numbers from 1 to 49 are chosen, and there are 13,983,816 ways of doing it (combinations of 49 elements taken 6 by 6), of which only one is the winner.

Supposing that this person plays 100 times a year, they will play 5,000 times during their life, the problem is similar to the birthday one but it is as if we had a year with 13,983,816 days and 5,000 people, each of which was born on one of those days. What is the probability that two were born on the same day? Applying the formulae we have seen (a spreadsheet is essential) the probability is 59%. Then it is not strange, if we do our sums properly, to find out that the same combination has come up twice.

Adjacent numbers in the lottery

We are going to finish this chapter off with a question that you may have asked yourself at some point. What is the probability that in the lottery draw 2 adjacent numbers come up.

It is higher than it may at first seem, exactly 49.5%. Calculating this value with the combinatorial formulas is not quick, but we can verify that it is in that order of magnitude using a spreadsheet.

One possible way is the following:

1. Place the numbers from 1 to 49 in column A.

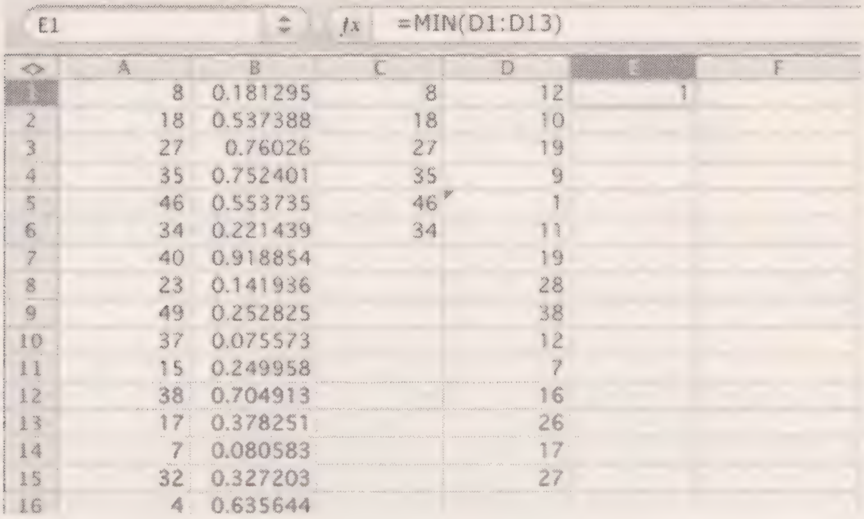
2. Put random numbers in column B.

3. Sort column B, including column A in the sort.

4. Column A is now sorted at random. Copy the first 6 values to column C. This will be the winning combination.

5. In column D enter the absolute value of the 15 differences between the values of the winning combination.

6. In cell E1 enter the minimum value of column D. If this value is 1 it means that there are adjacent numbers in the winning combination.



=MIN(D1:D13)

	A	B	C	D	E	F
1	8	0.181295	8	12	1	
2	18	0.537388	18	10		
3	27	0.76026	27	19		
4	35	0.752401	35	9		
5	46	0.553735	46	1		
6	34	0.221439	34	11		
7	40	0.918854		19		
8	23	0.141936		28		
9	49	0.252825		38		
10	37	0.075573		12		
11	15	0.249958		7		
12	38	0.704913		16		
13	17	0.378251		26		
14	7	0.080583		17		
15	32	0.327203		27		
16	4	0.635644				

Once this process is complete, if you resort column B, including column A in the sort, you will get another winning combination and the numbers will be re-calculated. The great thing about this is that you can keep pressing the F4 key and the whole process will repeat itself. You will see that approximately half the time a 1 appears in cell E1.

If you know any programming languages you could also write a small program that simulates the draw and count the number of times adjacent numbers come up.

Another option is to consult historical data. In the Spanish lottery from the first draw on 17 October 1985, up until 31 December 2009, 2,245 draws had been made. 1,148 of them (50.14%) had adjacent numbers.

And finally: on 22 August 2002 the winning combination was: 13, 21, 24, 26, 32 and 34, and on 10 December, 2009, it was… exactly the same! It is not that strange, the probability of a repetition in 2,245 draws is 16.5%.

Chapter 3

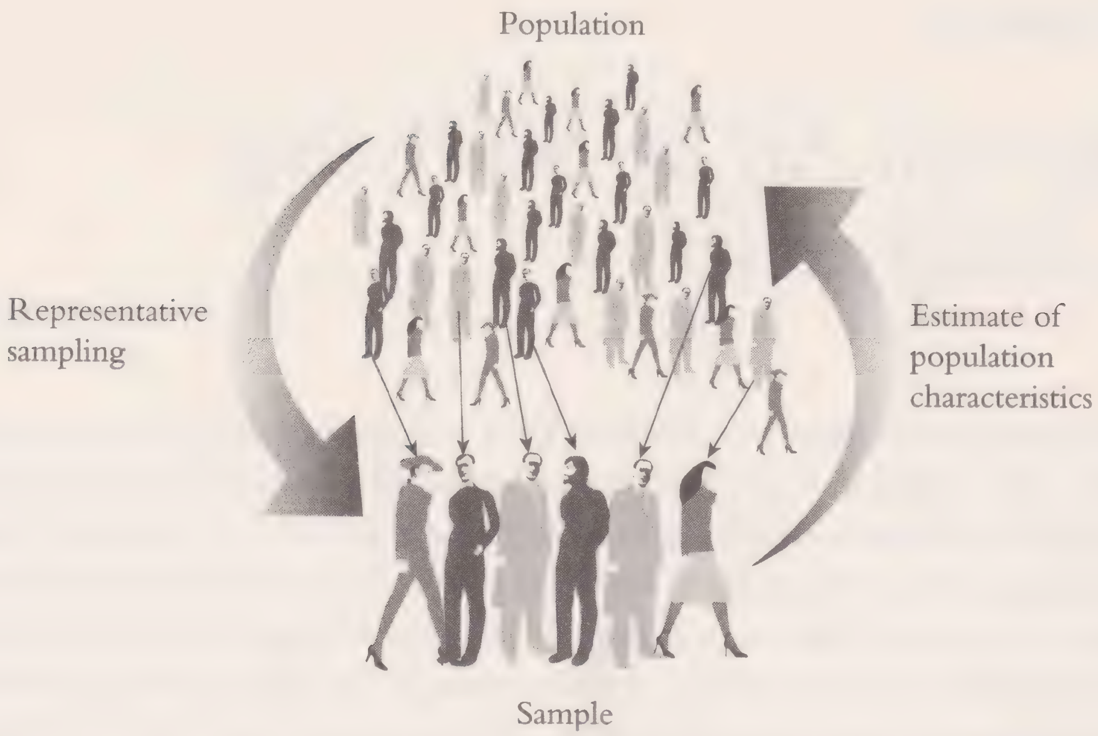
Reveal All by Studying a Fragment

One of the more typical activities in statistics is to draw conclusions on a whole while only looking at a part. The 'whole' is called the 'population', an inheritance from the first applications of statistics, in which the object of study was exactly that, a population of individuals. Now we still give it the same name, but the population does not necessarily have to be formed by people; they could be fish in a lake or the products of a factory during a year. Of course, it could also be made up by the set of citizens with the right to vote in the coming elections or by the set of people who suffer from a particular illness.

Studying the population exhaustively is nearly always impossible. It is not viable to ask the entire electorate who they are thinking of voting for in the next elections, or all sick people how they responded to a new drug. Similarly, testing machines and other products can involve measuring how much use or blunt force they can withstand before they break. Breaking every one of them to ascertain their precise durability is surely not a good idea.

What we do is choose part of the population, called the 'sample' and, from the results obtained for that sample, estimate (get the closest idea possible) the characteristics of the population. The rules for calculating probabilities allow us to quantify the quality of these estimates through a series of concepts such as 'level of confidence' and 'margin of error'.

Of course all of this will be true as long as the sample is representative of the whole population. If it is not representative, it will obviously be of no use, although sometimes in certain reports the importance of the mathematical aspects is exaggerated (as bamboozling people with mathematical language is a cheap and effective way to get agreement) and the way in which the sample was collected is overlooked. To do it properly is much more expensive, but it is absolutely critical for guaranteeing the validity of the conclusions.



Estimate of the characteristics of the population from a representative sample.

How many fish are there in a lake? How many taxis are there in a city?

Let's take a look at two examples of estimating characteristics of a population, in this case its size, by applying sampling techniques.

Fish

Counting how many fish there are in a lake seems like a difficult task, especially if the lake is large and the water dark, but biologists know how to do it. Using statistical techniques, of course. A method that is very commonly used is called 'fish and refish' (or, in general 'capture-recapture', because it is not only used for fish). The procedure is as follows:

1. Catch a sample of fish, mark them and return them to the water. Naturally this cannot be done in any old way. The fishing must be done in a way that does not hurt the fish. There are techniques such as the use of electric charges which stun them for long enough to be able to collect them and mark them. The mark must not obstruct the movement or survival of the fish and it also has to last until at least the next fishing trip.

2. Wait a while (perhaps a few days) until it is reasonable to assume that the marked fish have spread out across the whole lake, and return to fish another sample (the 'refish') which does not necessarily have to contain the same number as the original fish.

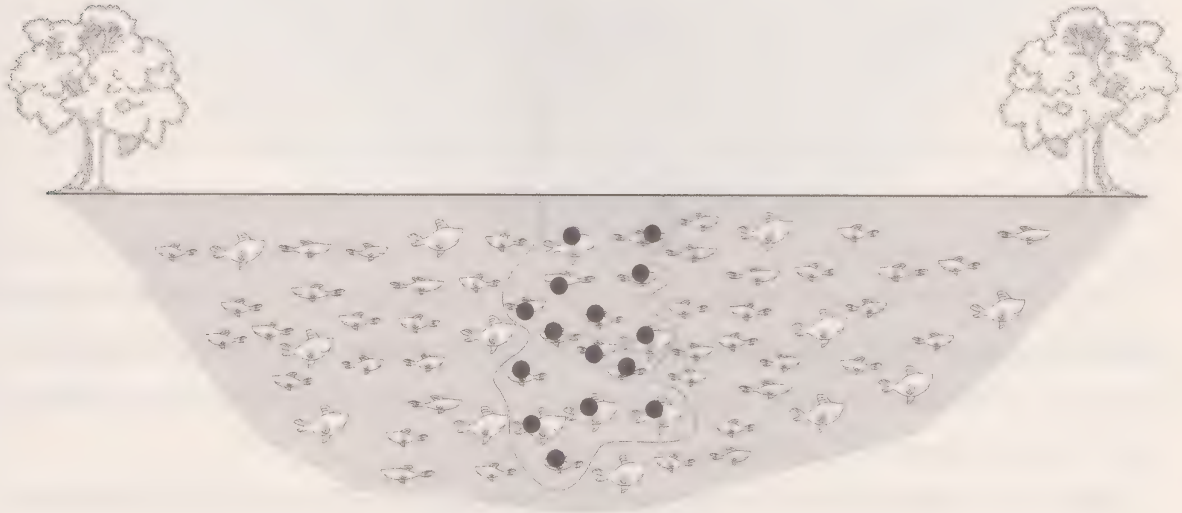
3. Do the calculations. If there are N fish in a lake and M are marked, the proportion of marked fish is M/N. In the refish C fish are caught, which can be considered a representative sample of all the fish in the lake, and among them R are marked. It is reasonable to assume that the proportion of marked fish in the second sample is similar to the proportion of fish marked in the lake, in other words:

$$\frac{M}{N} \cong \frac{R}{C}.$$

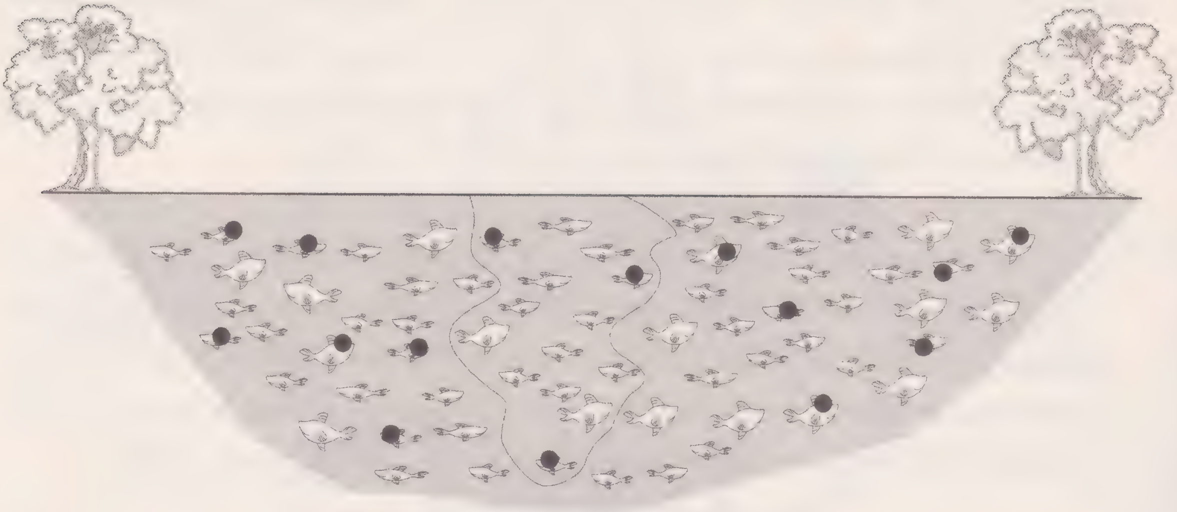
So that an estimate of the number of fish in the lake is (solving for N):

$$N \cong \frac{M \cdot C}{R}.$$

Here is a numerical example. 1) A set of M fish is caught and marked (which can be considered a random sample of the N fish in the lake). In our case, $M = 15$.

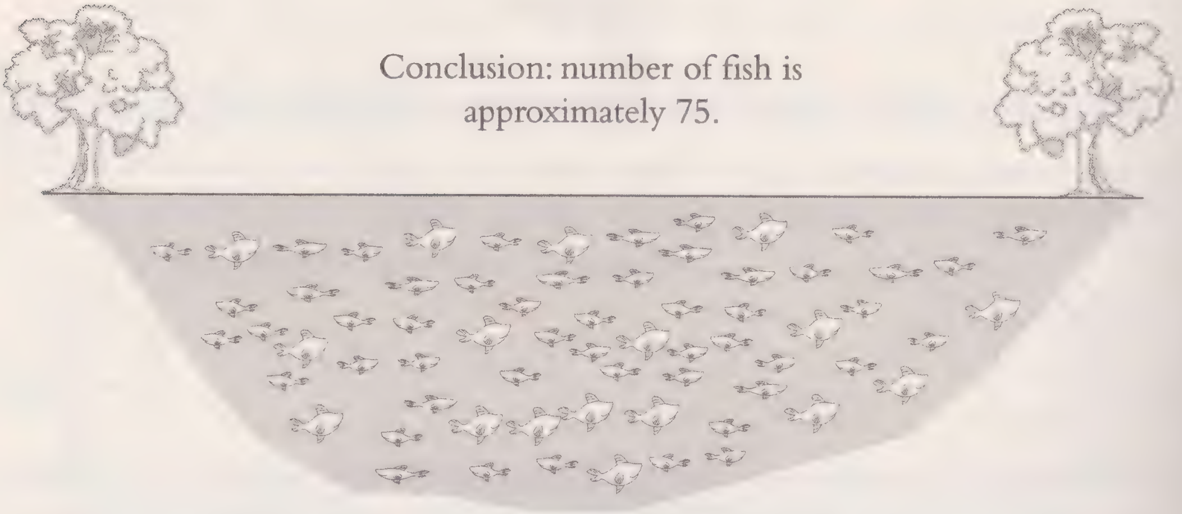


2. We wait a while so that the marked fish disperse around the entire lake and then catch some more (C) counting the number (R) of them which are marked. In our case, $C = 15$ and $R = 3$.



3. The number of fish in the lake will be around

$$N = \frac{M \cdot C}{R} = \frac{15 \cdot 15}{3} = 75.$$

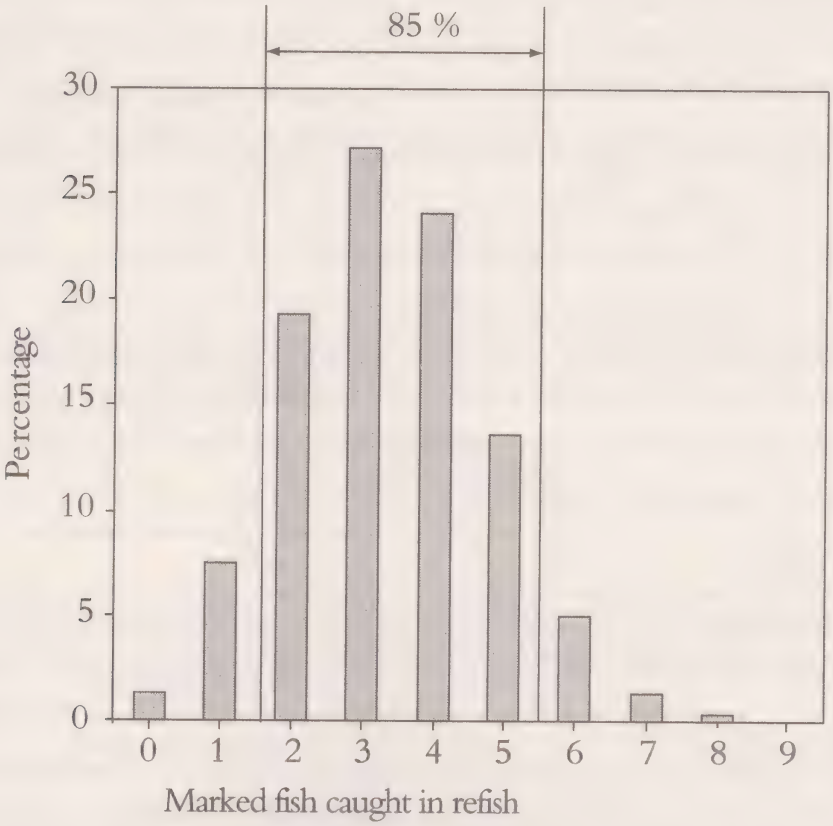


Conclusion: number of fish is approximately 75.

But what does 'approximately' mean? In the example, if you stop to count them, you will see that in the lake there are 67 fish and, therefore, there is an error of 12%. Is this error greater or lesser than what should be expected? What magnitude of error can be committed when applying this method?

Based on reasonable hypotheses and mathematical arguments, statistical theory can respond to these questions, although in order to get an idea we can also use a small computer program which simulates the process for a number of fish set by the user. Thus we can repeat the fish and refish as many times as we want, and, take the number of fish estimated in each exercise to determine the magnitude of the errors committed and how frequently each of them arises.

With the data from our example, in 85% of cases between 2 and 5 marked fish are found, which, applying the formula we have deduced, gives us an estimation of between 112 and 45 fish. 15% of the time the estimate is outside of this interval.



The distribution of the number of marked fish found in the refish with the data from the example. (Results obtained by simulating the process 10,000 times.)

Surplus errors are more frequent than deficit errors, and the mean value of the estimates is 82, higher than the real value. When this happens, we say that the estimator is 'biased', it does not point to the true value of the parameter that is being estimated.

The estimate notably improves by introducing small corrections to the formula. The only problem is that we can no longer justify it simply.

$$\dot{N} = \frac{(M+1)(C+1)}{R+1} - 1.$$

Using this formula, finding two fish leads to an estimate of 84, and if we find 5 the estimate will be 42. Therefore, in 85% of the cases the estimate will be between 42 and 84. Also, 27% of the time we will find 3 marked fish and our estimate will be 64, very close to the real value. This is an 'unbiased' estimator, in other words, if we repeat the program over and over, the average estimated value will coincide with the real value.

We can also include correction coefficients if we consider that not all the fish have the same probability of being caught, or if the added markers affect survival, or if some fish lose the markers. In short, this is a very well studied subject which is covered in detail by ecology books. It is also a good example of how statistics can allow us to overcome problems which at first may seem difficult, if not impossible.

Taxis

The challenge is much simpler if we are talking about finding the number of taxis in a city. The first option is to do a search on the Internet. For example, a large city council's web site states that in the metropolitan area there are 10,480 taxi licences. One vehicle corresponds to each licence. Case solved.

Number of licences	10,481
Taxi driver credentials	19,019 (12,536 active)
Daily runs	More than 225,000
Authorised taxi vehicles	13 makes / 55 models
Age	5 years: 67.5% Between 5 and 7 years: 19.6% Between 7 and 10 years: 9.5% + 10 years: 3.4%
Mean age of taxi fleet	4 years

But if we cannot find this data, we can use statistics. Taxis have a licence number that is visible in the car, and these numbers are correlative from 1 to the number of licences. When we buy a new car, the number plate is also new and our old car's number ceases to exist.

However, with taxi licence numbers it is different (although it is probable that there are exceptions): the number of licences is fixed, and if a person wants to be a taxi driver they have to buy the licence from another, and their number will be that of the licence which they have bought. This makes the task of counting much easier, and in 10 minutes, in the centre of the city, without telephone or an internet connection, you can make a very good estimate of how many taxis there are. Let's see how.

If the following values are taken from a numbered population: 8, 14, 22, 27 and 35, and you are asked for an estimate of the number of elements it contains, you would surely not say 25 because at a glance we can see that there are at least 35, and nor would you say 1,000 because it would be very strange to get 5 numbers at random that are so low if the numbers could go up to 1,000. Surely a good estimation would be around 40 to 50.

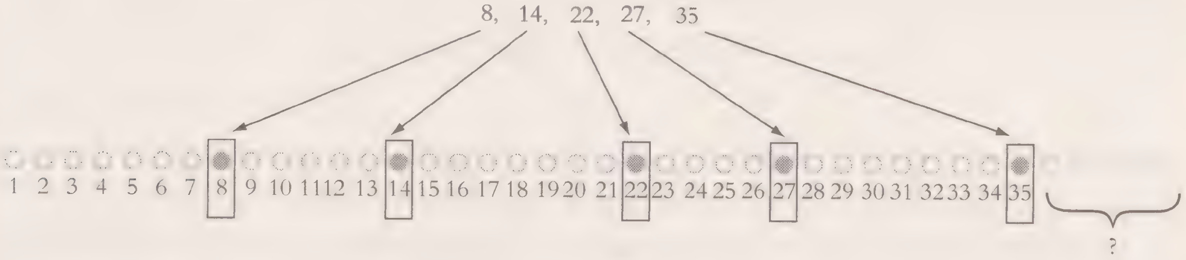
An initial idea for a rule could come from observing that in a population of this type the total number of items is equal to 2 times its mean minus one. For example, if the population consists of 10 elements, their values will be 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10, the mean 5.5 and the number of items in the population $2 \cdot 5.5 - 1$. In general if $\overline{x}$ is the mean of a population formed by N consecutive numbers starting from 1, it is always true that:

$$N = 2\overline{x} - 1.$$

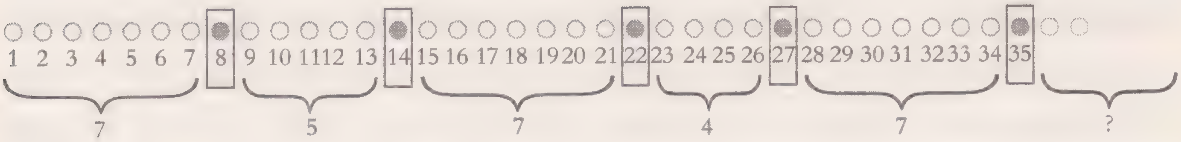
If we apply this general rule to the data from the previous sample, its mean is 21.2; then the estimate will be $2 \cdot 21.2 - 1 \cong 41$, which goes very well with what we intuitively thought.

But this system has a significant problem. If the data is 3, 4, 6 and 15, the average is 7 and our estimate for the total number of items is 13 which is obviously incorrect, since in the sample with have an item 15 and, therefore, there must be at least that number of elements. It is quite ridiculous, and not that unusual, to use sophisticated techniques to reach conclusions that common sense tells us are wrong. We will have to think of some new methods.

In reality we only need to know how many elements there are above 35 in order to know how many there are in total.



One option that seems quite logical is to assume that after the last item there are as many more as before the first one. In this case we would add 7 to 35 and our estimate would be 42. The inconvenience of this method is that we ignore the information contributed by the number of elements there are between the observations, and it is always good to make use of all the information available. One way of doing it is to add the last value to the average of the distances between the observations which we have (the first distance is the number of elements there are before the first observation).



In our case we will have to add:

$$\frac{7+5+7+4+7}{5}=6.$$

Therefore our estimate will be 41. In general, if x_1, x_2..., x_{n} represents the values in positions 1, 2..., n, the amount that should be added to the last value will be:

$$\frac{(x_1-1)+(x_2-x_1-1)+(x_3-x_2-1)+\ldots+(x_n-x_{n-1}-1)}{n}.$$

And it is easy to verify that this expression is equivalent to:

$$\frac{x_n}{n}-1.$$

Therefore, the best estimate of the total number of elements in a population is:

$$x_n+\frac{x_n}{n}-1.$$

And what is the quality of this estimator? It can be demonstrated (as mathematical statistics does) that with the criteria that are manipulated to qualify the characteristics of an estimator this is the best that can be calculated. In the jargon used by the experts this is called an UMVUE (*Uniformly Minimum-Variance Unbiased Estimator*).

Thus, it is sufficient to look at the licences of 20 taxis and the value of the highest number is added to that value and divided by 20, then we subtract one. In the case of our example, if the number of licences is 10,481 and they are correlatively numbered, in 95% of cases our estimate will be between 9,175 and 10,990.

Obviously, this method is not only useful for counting taxis. For example, it can also be used to estimate the number of participants in a race if the their numbers are handed out in order starting with 1 up to the last person to sign up. Also, completely changing the scenario, espionage uses these techniques to estimate how many weapons an enemy has. If the weapons have a serial number and a few samples are obtained, we have already seen that it is not difficult to deduce the total number at their disposal.

What proportion of homes have the Internet?

Firstly we need to be clear about the terminology: What is a home? What do we understand by Internet connection? It does not make sense to walk a fine line in the calculation of values before it is clear what they mean.

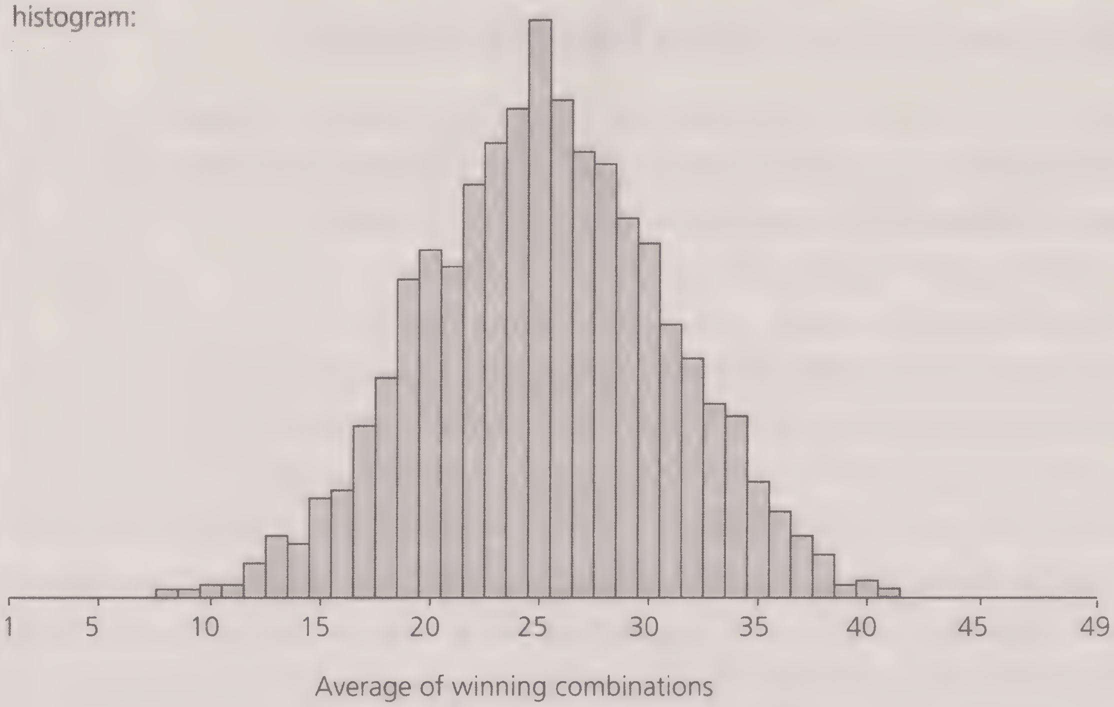
A newspaper headline said that half of all cigarettes are smoked by people with mental disorders, it sounds as if they are saying "half of all smokers are crazy" and that must be an exaggeration. But in the text of the article mental disorder is defined as the addiction to a substance, therefore not just half but practically all cigarettes are smoked by people with an addiction, and therefore, with "mental disorders". Many words that we commonly use have a confusing meanings and one of them is 'family'. What is a family? A marriage with kids? And what if the grandparents live in the same house should they be counted as members of the family? It seems strange that whether or not someone belongs to a family depends on the house they live in. Family can also be understood in a much broader sense, such as at weddings when we talk about the bride and groom's families, which can easily include several dozen members each.

Is a home the same as a house? Surely not, because if nobody lives there it cannot be a home. If it is weekend house, or it is only used during holidays, then surely it cannot be considered a home either. What about a student flat, which is only occupied during the academic year, is that a home? A home seems to be linked to a family, or is that not necessary? It is therefore essential to agree on what is understood by home.

ESTIMATES FOR THE WINNING COMBINATION FOR THE LOTTERY

We know all too well that all the numbers in a national lottery have the same probability of coming out, but what about the mean of the winning combination? On 7 January 2010 the winning combination in the Spanish Lotería Primitiva was 19, 24, 25, 38, 43 and 49, which has a mean of 33, and on Saturday 9 January it was 13, 26, 29, 30, 31 and 43, which has a mean of 28.67 (rounded up). Do all the means have the same probability or do some come up more often than others?

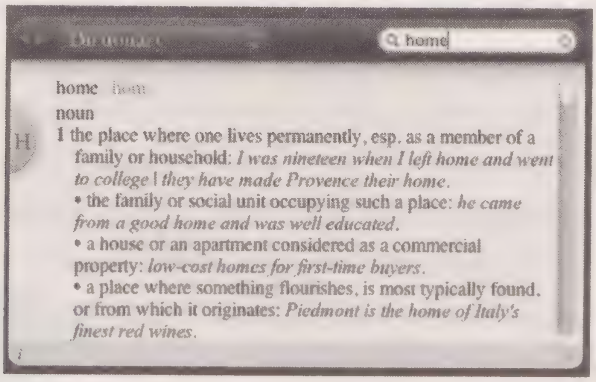
The answer is that some come up more often because, as we saw in the previous chapter, the means tend to follow the normal distribution pattern. The distribution of the mean of the draws made between 17 October 1985, and 31 December 2009, is shown in the following histogram:



It is much more probable that the mean is between 20 and 30 than between 5 and 15. So, why not always bet on the combinations that have a mean between 20 and 30? Because there are many more combinations that have those means, and the probability of getting the combination right is always the same.

In other words, if there are 1,000 numbers in a draw, which is more likely, that a number between 500 and 550 comes up or a number outside of that interval? Obviously, the latter is more probable, but that does not mean that a number within those limits has less chance of coming up than one which is outside them.

The meaning of having an Internet connection is obviously less confusing, it does not matter of it is through a modem or broadband. However, there are houses that have a wireless connection because their neighbour has it and it is not protected, or because they live near the library or in a free wi-fi zone; should these homes be counted as having an internet connection, or only the ones that pay for it?



home

noun

1 the place where one lives permanently, esp. as a member of a family or household: *I was nineteen when I left home and went to college | they have made Provence their home.*

• the family or social unit occupying such a place: *he came from a good home and was well educated.*

• a house or an apartment considered as a commercial property: *low-cost homes for first-time buyers.*

• a place where something flourishes, is most typically found, or from which it originates: *Piedmont is the home of Italy's finest red wines.*

Definition of 'home' in an English dictionary.

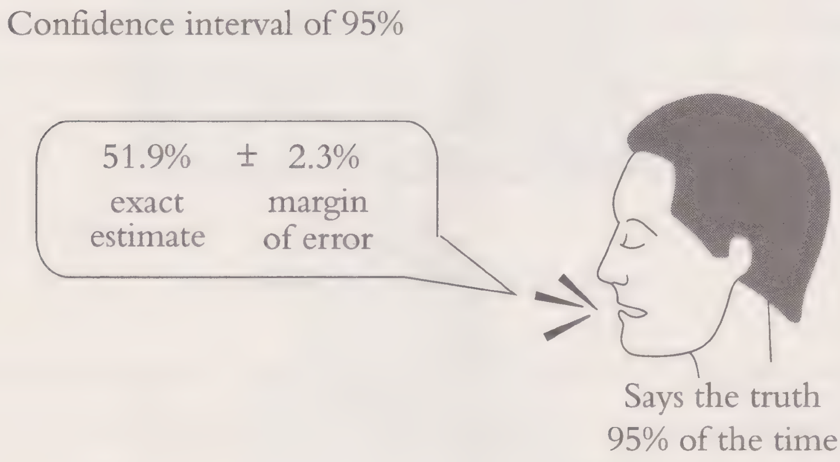
Let's assume that home is defined as a house in which one or more related people live for most of the year, and that it is understood that they have an Internet connection if it is actually paid for by them.

If we take a sample of 1,000 from a population of 100,000 homes and it turns out that the proportion of them that have an Internet connection in the sample is 51.9%, does that mean that this is the exact percentage for the whole population? Obviously, the answer is 'not necessarily'. If instead of the random sample which we have, we had got a different one, the result would undoubtedly have been different, it could have been 50.7% or 52.3% for example.

This is why, when the results of a study of this type are published not only the value of the estimated proportion is included, but also the reasonable margin of error around that value. For example, the result of the estimate could be: 51.9% ± 2.3%.

The 2.3% which we add or subtract is what we call the 'margin of error'. It means that although we have a specific value we cannot be sure that the true value is exactly the same. The calculation of probabilities allows us to determine the variability of our estimate and from it calculate the margin of error (this is a binomial distribution problem, here the experiment is to look at a home and the two possible results: it has an Internet connection or it does not).

The interval that includes the margin of error is called the 'confidence interval'. Can we be sure that the true value will be in this interval? The answer, once again, is no, we cannot be sure. The margin of error is calculated for a certain level of confidence and that level is often 95%, which means that it has been calculated by a procedure that we know works. It includes the true value of the sought proportion in 95% of cases, but we cannot be sure if it works in our specific case. It is as if a person who was right 95% of the time told us the interval; we can almost be sure that it is true, but not completely sure.



The concept of confidence intervals.

Confidence intervals of 99 or 99.9% can also be calculated but it is not often done because, given a sample size, the larger the level of confidence, the larger the margin of error which is obtained, and it is not of much use to say with great certainty that the proportion we are looking for is in the interval between 51.9 ± 40%, there is no need for a study in order to conclude that. If we want to increase the level of confidence while maintaining the margin of error the only thing we can do is increase the sample size (money solves a lot of problems, including this one).

"Party A leads Party B by 3.6 points"

Under a headline like the above it is common to find texts like the following in the press: "If the general elections were held today, Party A would beat Party B by 3.6 percentage points, according to the estimate of the probable vote carried out by company X. Three months ago, the advantage was 5 decimal points less. The data confirms a favourable trend for party A." And in a small box it then states, among other things, that the margin of error of the data for the whole sample is ±4.5%. A superficial analysis of this data shows that it is not so clear that A is ahead of B. If the survey gives a percentage of votes of 41.6% for party A, the margin of error indicates that a reasonable estimate for this proportion is between 37.1 and 46.1, and if party B starts at 38%, we are saying that the interval that we'd expect to find this value in is from 33.5 to 42.5. Therefore, it could also be, according to the data provided in the survey, that A receives 39% and B 40%.

What is certain is that if three months ago the advantage was 5 tenths less (in the survey results, not in reality!), this does not confirm, in any way, a favourable tendency for party A.

The million dollar question

The question that people who are carrying out a survey ask themselves most often has to be: What should the sample size be so that the results are reliable? The answer is 'it depends', and it depends on:

1. The accuracy we want our results to have or, in other words, the margin of error which we are prepared to assume. If we want the margin of error to be 1%, we need a larger sample size than if we accept one of 4%.

2. The certainty, called 'level of confidence', with which we wish to make the estimate. If we agree on an 80% confidence level we will need a sample size smaller than if the interval needs to be 95%.

3. The actual value of the proportion that we want to estimate. Although at first this may seem a little strange, it is logical that it should be so. If there is no variability in the population (100% of the items are equal) we only need one item to know everything. If all the balls in a pot are white, or all of them are black,

we only need to take one out in order to find out what colour they are. The greater variability there is, the greater the size of the sample required, and the most favourable case is when the proportion is 50%. What we do is assume a value for this proportion trying to err on the high side. If we do not know anything or want to be conservative, we assume that it is 50% and we can be sure that the sample size need not be any greater than what we get. If we know that the proportion we are looking for is lower (for example, the percentage of houses that have a fax machine), we could assume, playing it safe, that it is 20% (it is undoubtedly less).

4. The size of the population. When the population is small (let's say up to 100,000 individuals), and if the margin of error that we are looking for is also small (1 or 2%), the greater the size of the population the greater the sample size required. But for large populations, or margins of error of 5% or more, the influence of the size of the population is barely noticeable. There are quite a lot of misunderstandings on this subject, on which we will go into more detail later.

THE SIZE OF THE SAMPLE

In case you need it some time, the formula that links all the ingredients for determining the size of a sample is:

$$n = \frac{z^2_{\alpha/2} p\, q\, N}{E^2 N + z^2_{\alpha/2} p\, q}.$$

Where:

$z_{\alpha/2}$ is the value linked to the confidence level. If it is 95%, the most common case, then its value is 1.96. Sometimes it takes a value of 2, and in this case it corresponds to a confidence interval of 95.5%.

p is the proportion we are trying to estimate.

$q = 1 - p$.

E is the margin of error.

N is the size of the population.

Now we just need a spreadsheet in order to start testing and seeing what happens to the size of the sample when the level of confidence increases or the margin of error changes, or what effect any of the variables involved has. We can also make a table such as the following, in which it is almost all done for us.

Size of the population	Margin of error					
	±1%	±2%	±3%	±4%	±5%	±10%
500	467	414	341	273	218	81
1,000	906	706	517	376	278	88
1,500	1.298	924	624	429	306	91
2,000	1.656	1.092	696	462	323	92
2,500	1.984	1.225	748	485	333	93
3,000	2.286	1.334	788	501	341	94
3,500	2.566	1.425	818	513	347	94
4,000	2.824	1.501	843	522	351	94
4,500	3.065	1.566	863	530	354	95
5,000	3.289	1.623	880	536	357	95
6,000	3.693	1.715	906	546	362	95
7,000	4.049	1.788	926	553	365	95
8,000	4.365	1.847	942	559	367	95
9,000	4.647	1.896	954	563	369	96
10,000	4.899	1.937	965	567	370	96
15,000	5.856	2.070	997	578	375	96
20,000	6.489	2.144	1.014	583	377	96
25,000	6.939	2.191	1.024	587	379	96
50,000	8.057	2.291	1.045	594	382	96
100,000	8.763	2.345	1.056	597	383	96
500,000	9.423	2.390	1.065	600	384	97
1.000,000	9.513	2.396	1.066	600	384	97
1,500,000	9.543	2.398	1.067	600	385	97
2,000,000	9.558	2.399	1.067	601	385	97
50,000,000	9.602	2.401	1.068	601	385	97

Table with the sample sizes necessary for a level of confidence of 95% and in the least favourable case in which p = q = 0.5.

Surprise! The size of the sample does not depend on the size of the population

There are some ideas on sample sizes that, despite being quite well known, are completely false. For example, occasionally there are questions about the results of a survey that raise the argument: 'the sample is not representative, it does not even cover 10% of the population'. Figures such as 10%, like any other number, are completely arbitrary. Professor Roberto Behar, of the Universidad del Valle in Cali, Colombia, explains it with some clear analogies.

Does the soup need more salt?

To prepare soup we need a saucepan, let's say it is a small one, and in order to know if it is lacking salt we taste it with a spoon. If we have guests over and we need to prepare the soup in a much larger saucepan, do we also need a larger spoon to taste it? Obviously not. We all use the same spoon, and we sip it in the same way, both when the pan is big and when it is small. The size of the sample does not depend on the size of the population.

What we do need to do, whatever the size of the pan, is mix it well in order to homogenise the soup, and make sure that any potential sample provides the same information. Nobody is surprised that it is much more important to mix well than to increase the size of the spoon. And we know that the error of not stirring cannot be corrected with a larger spoon. If the sample is not representative, increasing the size does not solve the problem. At all.

What is my blood group?

A drop of blood is enough to unequivocally identify someone's blood group, given that every drop of blood from a person is of the same type. If you have seen one, you have seen them all. This again demonstrates that the impact of uniformity is much more significant than the size of the population. The same amount of blood is required for a newborn baby, weighing just a few pounds, as for its father, who weighs good deal more.

But the intuitive argument is not the only one. We can also use the formula to see what the relationship between the size of the sample and the size of the population is. If the population is small the sample grows rapidly as the population increases, but from a certain value it remains practically stable.

LEFT-HANDED PEOPLE DO NOT LIVE AS LONG (OR DO THEY?)

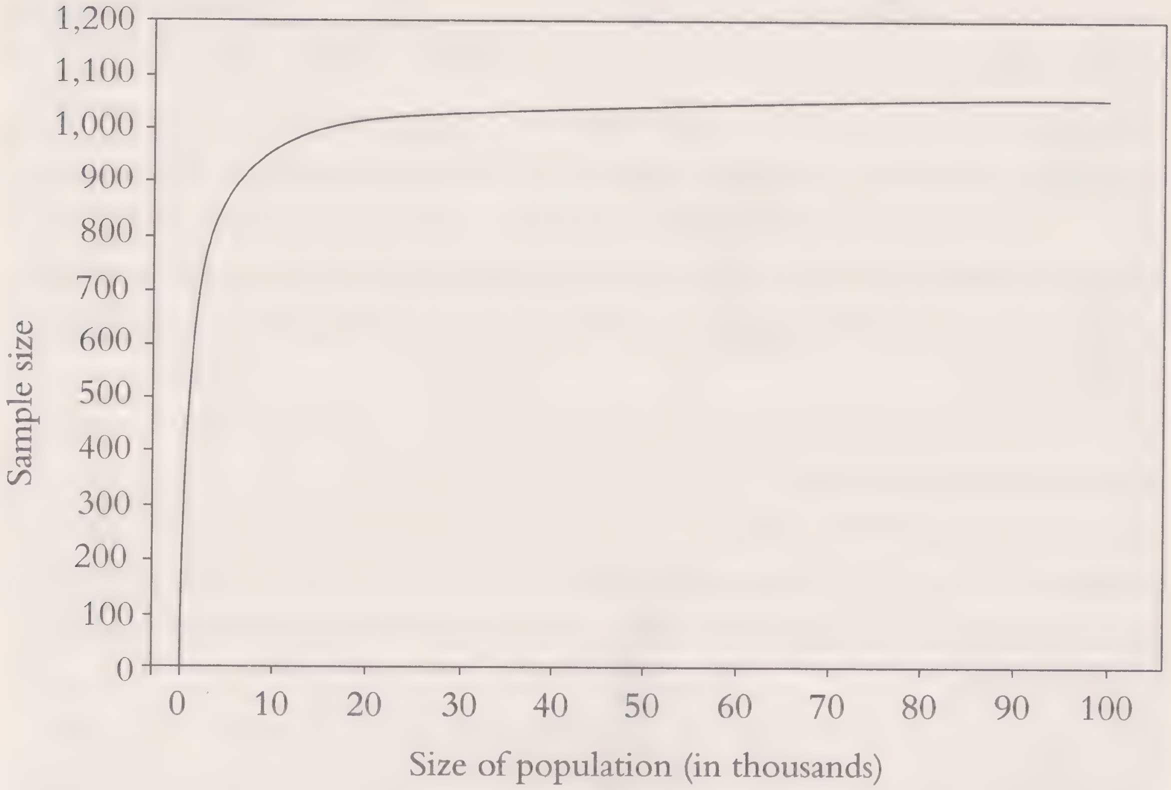
On 4 April 1991 the *Washington Post* published an article on its front page about a study which demonstrated that left-handed people live, on average, 9 years less than right-handed people. The study was based on analysis of deaths in two counties in California in which the age of death was analysed according to whether the person was left- or right-handed. While right-handed people frequently reached old age, it was far less common among left-handed people.

The news had a great impact and explanations for the results quickly came about: it was said that left-handed people are more prone to certain types of illnesses and to suffering serious accidents; one of the causes could be that machines, the devices which we use from day to day, the world in general, are designed and made for right-handed people. For left-handers this caused maladjustment, accidents, etc. and as a consequence of all this a considerable decrease in life expectancy.

But no, in February 1993, in the *American Journal of Public Health* a rigorous and well documented article was published which put things in their place: the difference in the age of death could be explained completely by the difference in the distribution of ages of right- and left-handers. At the beginning of the 20th century when a child showed a tendency to eat or write with their left hand they were forced to do so with the right, so that when the study was carried out there were very few old left-handed people and, therefore, very few people of that age died, not because they did not reach old age, but because they were not allowed to be left-handed.

This article did not appear on the front pages, confirming that the most surprising and spectacular stories always have more impact. This case also demonstrates that with a little data it is easy to find believable reasons to justify them. I think that people who give reasons for why the stock market rose or fell the previous day know a little about this.

For a margin of error of 3% and a confidence level of 95% in a population of some 10,000 individuals, a sample of about 1,000 individuals is needed. From this value the required sample size grows very little. For a population of 100,000 a sample of 1,056 is needed, for 1 million the sample size has to be 1,066, and for 50 million, 1,068. We need the same sample size for a small city as for an entire nation.



The relation between the population size and the size of the sample for a margin of error of 3% and a confidence level of 95%.

But it is necessary to ensure that the sample is representative. If that can be guaranteed, if the soup is well stirred, the size of the spoon does not matter.

The power of randomness

Sometimes reports on the results of a survey highlight the calculations of the levels of confidence, but they barely touch on the way in which the sample was obtained, or they explain it and it is clear that the sample is not random. All the mathematics behind these calculations is based on a few conditions which are only met when the sample is random. If that is not the case, it is being attributed a significance that it does not have, and the level of confidence will be incorrect, however well the calculations are done.

The best way of selecting a random sample is to have a list of all the individuals in the population, select the sample at random and go after them: locate them, agree a time and date for the interview, meet up, etc. The problem is that it is very expensive. Another option is to select houses; it is easier, but during the day people who work are not at home and at night they rarely want to deal with interviewers. Also, if you have to wait for the night to do the interview, you will get little done in a day.

Completely random samples have the advantage that the statistical estimation methods work very well, but as we have already said they have the disadvantage of being expensive. There are other variants, each one with its pros and cons, such as stratified sampling (the population is divided into layers and the samples are taken from those layers, which is more efficient if there is little variability) and conglomerate sampling (for example, instead of selecting individuals, blocks of flats are selected and everyone who lives there is interviewed, which is cheaper than using individuals who are spread out geographically). The companies that work in this field know how to do it in order to achieve the necessary levels of reliability in a way that is economically viable. But it is always critical to ensure that the sample is representative. Being careless at this stage can lead to infamous failures.

The poll that changed polls: Landon against Roosevelt

In the 1936 US presidential elections Alf Landon was the Republican candidate and Franklin D. Roosevelt the Democratic candidate. The *Literary Digest*, a respected and influential magazine from that period, which had successfully made predictions in previous electoral contests, carried out the largest electoral poll in history. It sent about 10 million questionnaires by mail taking the addresses from the lists of car owners and telephone directories. 2,300,000 responses were received and from them it was deduced that Landon would win with a proportion of three votes to each of Roosevelt's two.

As you may have guessed, Roosevelt won, and what is more, he did it by a wide margin (60.8% of the votes). The error came about because the sample from which the results were estimated was not representative of the general voters. In 1936 cars and telephones were only available to the moneyed classes (who tended to vote Republican). The effort was huge, and the failure, even more so.

In contrast with that study, a company more recently created by George Gallup correctly estimated the result by consulting fewer than 5,000 people, but ensuring that the sample was representative. The lesson was learnt: polls would never again be done by 'brute force', and Mr Gallup's company became a benchmark for fairness in opinon polls.

Controversial draws for military service

When seeking to obtain a sample, or just a number, at random, the details need to be taken care of because otherwise unforeseen problems can arise. A case which is often cited by specialists is that which occurred in the US when a draw was made in order to determine call ups to military service in the Vietnam War.

It was the first time that a draw of this type had been made and in each box 366 capsules were placed, each one with the date of a day of the year. First the 31 for the month of January were put in, then the 29 from February and so on up to the 31 capsules for December. They were mixed up and names began to drawn out. Those who were born on the date that was drawn first were the first to be called up, then those born on the date taken out second, and so on until the end.

The problem came about because, it appears, the capsules were not mixed up properly. The dates in December, which were the last to be added, stayed on top and and came out first in a proportion which was too great to be random, while those from January remained at the bottom and came out towards the end, so that males born in December were recruited and sent to Vietnam in a greater number than those born in January. The media realised the problem and denounced the results, but they remained unchanged. However the system was changed the following year and the draw was made in a truly random way.

In Europe, more specifically Spain, a similar thing happened. In 1997 there were 165,342 youths at the call-up age for military service, but there was not enough space for all of them. There were 16,442 too many, so a draw was made to decide who would be excluded from the draft. A number was assigned to each man and the idea was to take one out at random and the person with that number and the following 16,441 would be free. The problem came from the way in which a number was chosen at random from the 165,342.

First a number was taken out of a drum that contained the 0 and the 1 in order to decide if the number would be between 1 and 99,999 (when a zero was drawn) or between 100,000 and 165,342 (where a 1 was drawn) and 1 came up. Next, a number from 1 to 9 was drawn from a second drum and the 8 came out. As the number would have been just over one hundred and eighty thousand, greater than they wanted, another ball was taken out until a number lower than or equal to 6 came up. Any problems there? Yes, the probability of a number between 1 and 99,999 was the same as that between 100,000 and 165,342, but in the first case there are more values than in the second, and that means that for some the probability of being released was 8.2% while for the others it was 12.6%, more than 50% greater.

Informal surveys

A professional body sends a letter to its members asking them to fill in a questionnaire about their work and annual income. The objective is to create a report which is useful for the members themselves as a reference for negotiating their salaries. They are asked about the type of company they work in – multinational, family, large, small, with a long history, recently created, etc – about the sector, their position, how long they have filled the position – at the company, in the profession, etc. – and, finally, what their fixed gross salary is and the bonuses they commonly receive. The letter includes a stamped envelope for returning the questionnaire by post. 357 members responded from a total of 5,000 letters sent, the conclusions are a level of confidence of 95% and a margin of error of 5%.

Even if we refer to the table of sample sizes we can see that the numbers fit, the problem is that the sample is not random and, therefore, none of the values make sense. Self-selected samples (we ask everyone and anyone who wants to replies) can never be considered random. It is possible that those working in executive positions are very busy, they travel a lot and have no time to reply to this type of questionnaire, or those who arrive home late, or those who earn very little or are on the dole and do not feel like dwelling on the subject, or those who have a salary structure that does not fit with that provided in the questionnaire are put off. In short, it is not a random sample, and in that case, the mathematical deductions which are based on those samples cannot be applied.

The same thing happens with questionnaires that we sometimes find in hotel rooms looking for our opinions on the facilities or the quality of the service. Undoubtedly only those who are particularly dissatisfied and find an outlet in the questionnaire, or those who are thankful for something and they want to put it into writing (and maybe those who have too much time on their hands and they spend it filling it in) will respond. The information which is collected may be useful for identifying things that have been done well and that have been done badly, but not for collecting reliable statistics on the opinions of the guests who have stayed in that room.

Going out on the street, microphone in hand (and camera on shoulder) in order to find out what people think about a controversial issue and then broadcasting their opinions on television after a sentence like "we went out on the street to find out what Londoners think about..." may make the programme more dynamic and entertaining, but it is no use for finding out the citizens' true opinions.

Yes or yes? The influence of how questions are phrased

The way in which the questions are worded, the order in which they are asked or the emphasis placed on particular words may affect the answers which are given. If a 'correct' answer is implied, the interviewee will tend to respond with what they think the interviewer wants to hear.

When this author and a colleague, who also works in statistics, ran summer courses on our subject, we used to demonstrate how the way in which the question is asked influences the answer by giving the attendees a questionnaire. We explained that we were looking for their opinion on a prospective new law on financing for political parties, and we handed out some sheets of paper which all looked the same, but on one half the question was worded in one way and on the other half, in another.

Do you think that there should be a law that prevents big financial groups from contributing large sums of money to election campaigns?

□ YES □ NO

Do you think that businesses and organisations should be allowed to donate, in a controlled and transparent way, to the parties that they support in electoral campaigns?

□ YES □ NO

Two ways of asking a question on the financing of political parties.

Nearly all the participants answered yes, regardless of which question they received. But some said yes to "preventing big financial groups from contributing large sums of money" and the others to "businessmen and organisations can contribute funds". As you will have noticed, depending on what you want the answer to be, you can ask the question in a different way and the problem is solved. But what the question is and the way in which it is asked are equally important, and also the exact questions should be published along with the results.

The phone rings… but you're not home: telephone surveys

The easiest and most comfortable types of survey are telephone surveys, but they also have their obvious inconveniences. There is a telephone available to almost everyone, at least in technically developed areas, but a new problem could be that young families only use mobile telephones, therefore their number does not appear in telephone books and they cannot be selected for this procedure.

It has to be considered whether the fact that houses without a land line have not been surveyed can affect the answer. The time of the phone call, who we ask to speak to and how to substitute those who choose not to answer also carry great importance. A lack of care at this stage can lead to serious errors in the predictions due to the sample not being representative.

A specific case: election polls

Election polls are one of the most talked about applications of statistics (and not always in a good way). These types of studies are unique for the amount of interest they attract and because, unlike other cases, eventually we always find out the true value of the parameters which have been estimated (the result of the election). The problem is that, as well as the habitual difficulties in finding random samples, there are other specific difficulties. Let's take a look at some.

Changing voting intentions

The results issued are based on surveys carried out several days, or even weeks, before the elections. In some countries it is illegal to publish election poll results during a certain period of time before the elections.

Thus, there are two types of extrapolation: that which is made from the sample of the population, this being the one subjected to statistical sampling theory, and that which generalises the results from the dates on which polls have been taken, projecting results for the day of the election.

But the parties are very busy on their election campaigns. There are debates between the candidates and events can occur that change their positions, and all this can affect the intention to vote, especially of those who were still unsure who to vote for at the time of the survey.

Who will the undecided vote for?

The undecided pose a big problem for those in charge of carrying out election polls. It is not unusual for the percentage of those who still do not know who they will vote for to be between 20 and 50% of the respondents. In these cases the votes are assigned based on answers to questions such as "Which party attracts you more" and "Which party do you feel closer to?" and also "Which party did you vote for in the last elections?"We are talking about an expert 'guessing' which party someone will vote for, even though they do not yet know themselves.

It is clear that assigning undecided votes to one party or the other is a task of critical importance, and its success depends more on knowledge of sociology and politics that it does on statistics.

HOW TO OBTAIN CONFIDENTIAL INFORMATION WITHOUT THE INTERVIEWEE FINDING OUT

When dealing with socially unacceptable or very individual behaviour, it is very easy to be deceived by the interviewee. But there are ways of obtaining that information while preserving the respondent's privacy, even in the presence of the interviewer. For example, let's assume it is embarrassing to answer "yes". So that the interviewee can speak without fear, we could do the following:

1. We ask them to choose a card from a pack, in which half are red and half are black. Only he looks at the card, and then he returns it to the pack.
2. If he picked a red card he responds "yes", if he chose black he responds to the question he is asked.

It is clear that if he responds "yes" the interviewer has no way of knowing if he has a red card or if he is responding to the question, which guarantees confidentiality.
If 1,000 interviews are done and 612 people respond "yes", approximately 500 will do so because they chose a red card and those answers can be disregarded. Of the other 500, those who really have responded to the question, 112 have responded positively, then our estimation is 112/500 = 22.4%.

Lack of sincerity in responses

The writing of the questions to be asked and the order in which they are written are also critical aspects. Writing clear questions, which do not tend to give predetermined answers, is not an easy task and it requires good knowledge of question asking technique and also requires having well trained and highly motivated interviewers (read: well paid).

Sometimes there are conditions of individual freedom of expression, which make the citizens' responses more or less believable and which will make the volume of the so-called 'undecided' greater or lesser, as it maybe that in place of 'undecided' what they actually have is 'guarded decisions', which they prefer to keep private.

From the percentage of votes to the number of seats

On many occasions what is truly relevant, more so than the percentage of votes which each party will receive, is the number of seats, and the systems that are used to distribute the seats based on the percentage of votes have complicated matters. For example, take a specific electoral situation in which there are 5 seats at stake and we can predict with 95% confidence that a certain party will obtain 32% of the votes with a margin of error of 3%. The problem is that if it gets 31% the party will get one seat, whereas if it gets 33% it will get 2. And that is an important difference but one that cannot be distinguished from the information that is available.

Another problem is that some legislations demand a minimum percentage of votes (for example 5%) in order to be included in the allocation. If a party is bordering on that percentage (for example, if it is estimated that it will get 4.6% of the votes) we cannot know if it will be included or not, and whether the outcome will also affect the number of seats for the rest of the parties.

But statistics do work!

When election polls are made there are many difficulties in achieving good predictions, difficulties which go beyond those involved in statistical sampling theory (not to mention manipulation and influenced results). It would be convenient to have a measure of the frequency and magnitude with which serious election polls fail (we won't mention those that are not serious), because just as it is always the bad news which is imposed on us by the media, more attention is always paid to gaffes in the predictions than when they are correct. Even in the academic environment, it is more sensational, sometimes more educational and always more welcome, to illustrate the way it should not be done than giving examples of when the predictions have worked well.

There can also be, and in fact there are, surveys that are produced by an interested group which is trying to influence the opinion of the electorate. The experience and seriousness of the company responsible for the study, as well as the media in which it is published, are also good indicators of the confidence which the polls deserve over and above the 95% which is normally indicated in the key.

Chapter 4

How We Make Decisions: Contrasting Hypotheses

It was the end of the 1920s in Cambridge, England. A group of professors, their wives and some guests were having tea outside, making the most of a nice afternoon. With the cup in her hand and after taking her first sip, a lady commented that it tasted different if the tea was added before or after the milk.

With the utmost politeness, of course, someone made some comments on how impossible that was, and so the discussion began, delving into all types of arguments based on physics and chemistry: the composition of the resulting product was the same both if the tea or the milk are added first, the dissolved particles ended up being exactly the same, the gradient of temperatures has no influence, etc. It was not possible to tell one cup from another... or perhaps they were missing something?

One of the people present, a 40-year-old man called Ronald Aylmer Fisher, proposed a 'revolutionary' process to clear up the matter: a test. Of course, it could not only be done with one cup of each type, because the probability of getting it right would be a half, and if she got it right they would not know if it had been luck or whether she really had been able to tell the difference between the two. But if they gave her 4 cups of each type, the probability of her getting it right was 1 in 70 (there were 70 different ways of choosing 4 objects from 8); so if she got it right in these conditions they could prove that she could tell the difference between one method of preparation and the other with a probability of error that was known to be small.

Fisher was already a famous professor at that time, and in 1935 would publish a benchmark text that marked the turning point in strategies for collecting data through experimentation. The book is called *The Design of Experiments,* and in Chapter 2 it introduces some of the key concepts using this case as a common thread.

The reasoning behind the tea taste test

First let's assume that the tea taster does not know how to tell one from the other; this is what we think most probable. We only believe in her ability if the data collected from a well-designed and controlled experiment go against the initial hypothesis. 'Going against' means that the results are not very likely in the case the she really cannot tell the difference, and we ourselves shall decide what not likely means: if it occurs less than 5% of the time, less than 1%, or any other value.

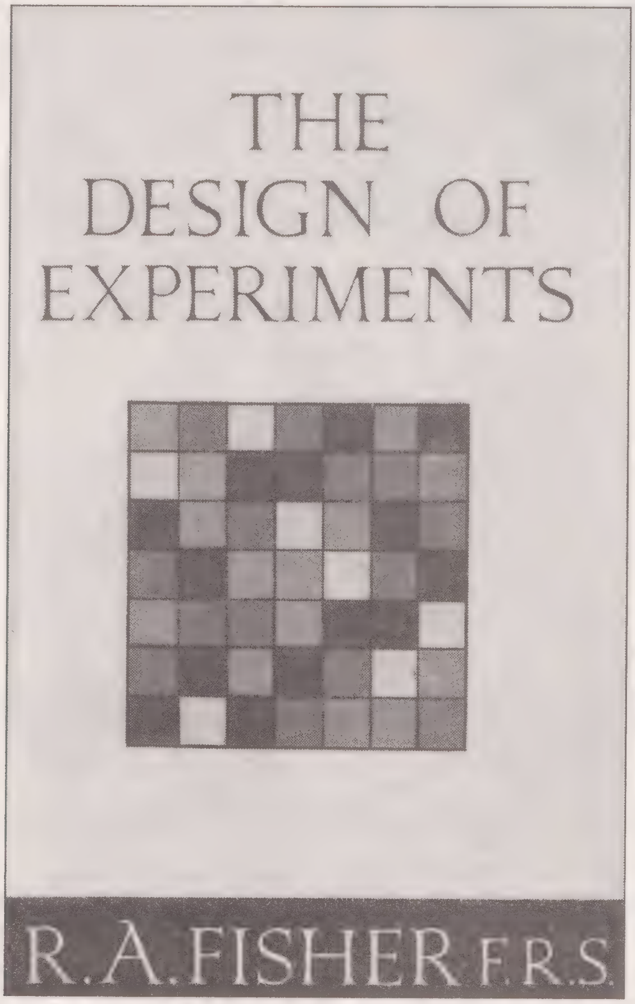
If we are only willing to believe her when the result of the experiment occurs at random (by coincidence, where the lady has no effect) less than 5% of the time, an experiment in which she is given 3 cups of each type would be of no use, as there are 20 ways of selecting 3 objects from 6 and only one is correct; then the probability of getting it right randomly is 1 in 20, or 5%. It is not difficult to work this out: the first mug can be chosen from among 6; the second from 5 and the third from 4, therefore we have $6 \cdot 5 \cdot 4 = 120$ ways of choosing 3 mugs. However, in this calculation we have taken into account the order in which they were chosen, in other words, supposing that we have labelled the mugs A to F, we have counted selecting ADF as different to choosing FDA. In order to subtract the repeated cases we have to divide by the number of orders possible with 3 cups ($3 \cdot 2 \cdot 1 = 6$). Therefore, the number of ways of choosing 3 mugs from a group of 6 equals $120 / 6 = 20$. If we have 4 of each type the number of ways of selecting the 4 will be: $8 \cdot 7 \cdot 6 \cdot 5 / 4 \cdot 3 \cdot 2 \cdot 1 = 70$, and as there is only one set of 4 in which the tea was added before the milk, the probability of getting this set right at random is 1 in 70, or, 1.4%. If the taster makes a mistake with one of the four cups selected, it would no longer be reasonable to consider that she can tell the difference, as the probability of this occurring is nearly 23%.

But we should not focus all our energy on mathematical reasoning. We also have to pay great attention to the details of carrying out the experiment and not give the taster clues... Fisher describes this, insisting that the cups are presented in a random order:

> "Our experiment will consist in preparing eight teas with milk, four in one way and the others in the other, and taking the mugs (in a randomly selected order) to the taster so that she may give her opinion. It will already have been explained to her what the test will consist of. She will be given eight mugs to try, four of each type, in a randomly selected order (by dice, roulette, cards, etc..., or simply by numbers published in some way). Her task consists of separating the mugs into two groups of

four sorting them, if she can, according to whether the tea or milk was added first.

And what happened? Fisher did not give the result of the test in his book, but among those present was professor Hugh Smith, who told the story to David Salsburg, the author of an excellent book on the explosion of statistics in the 20th century, *The Lady Tasting Tea.* The text begins by telling this story which, of course, gave the book its title. And he says that Professor Smith told him that the lady identified each and every one of the mugs correctly.



The Design of Experiments, *a classic and pioneering book in its field, in which Roland Fisher uses the example of the lady tasting tea to illustrate the key ideas in statistically significant tests.*

RONALD AYLMER FISHER: THE RIGHT PERSON AT THE RIGHT TIME

Born in 1890, Fisher was a scientist with solid mathematical training whose significant contributions in the fields of statistics and genetics set him apart. Although there are no official rankings he is undoubtedly one of the people, if not the one, who contributed most to statistics in the 20th century.

According to some sources, when he was young he was a frail child but was very keen to learn and very interested in astronomy. He also had serious problems with his sight, therefore doctors prohibited him from reading with artificial light (which was not the same as the lights we use today). This reduced his opportunities to study and, so that he did not get behind, he had a professor who taught him mathematics without using paper, a pencil or any type of visual aid, which awoke a deep geometric imagination that would later allow him to take on and resolve difficult problems from a highly original geometric viewpoint.

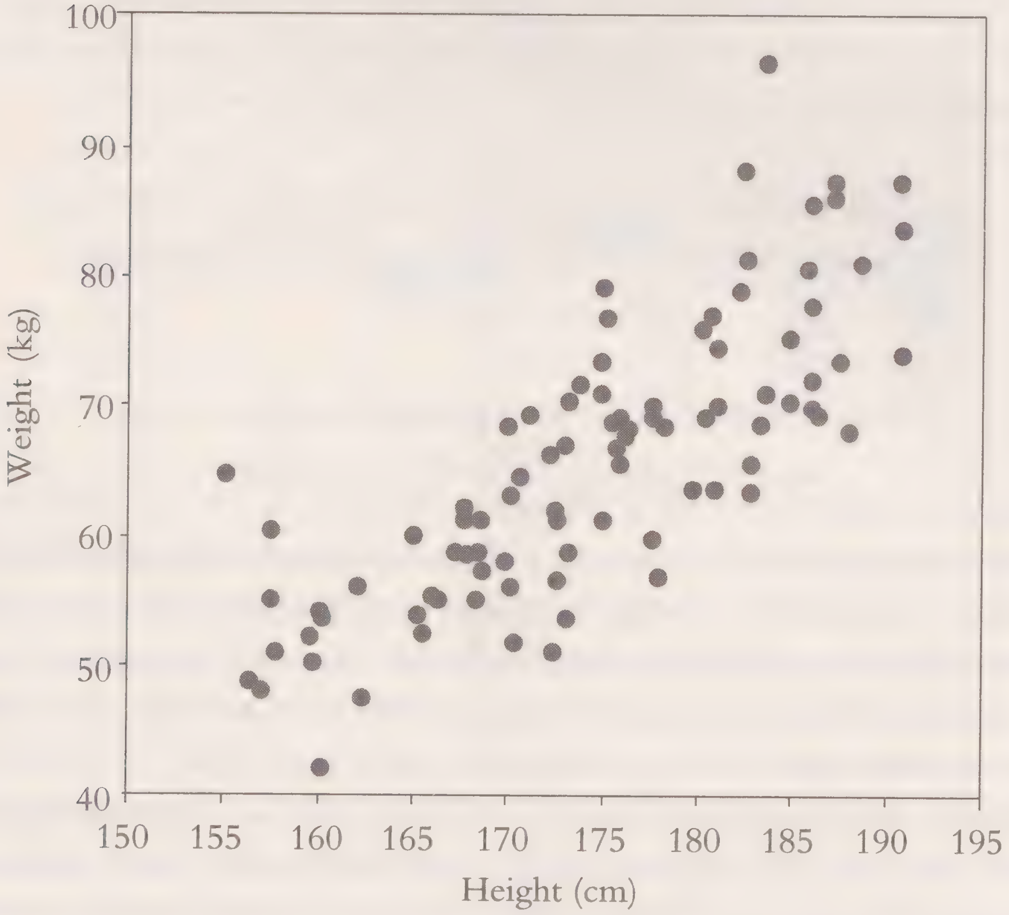
At 29, he and his wife, who was then 20 years old and with whom he had 3 children (other things have changed, apart from electric lighting), he moved to an old farm near the Rothamsted experimental centre just north of London. He was hired by the centre's owners, who made fertiliser, to organise the enormous quantity of data that had been collected during the centre's 90 years of operation.

Fisher demonstrated that, due to the way the data had been collected, the influence of rain and the weather in general were masking the possible influence of the fertilisers that were being tested. In current terminology we would say that both factors were 'confused'. But not only did he say that it was being done badly, he explained how it should be done and published a book, *The Design of Experiments*, which marked the beginning of a new era in experimental plans and data collection that has had great repercussions in agricultural and industrial research.

Weight, height; the coefficient of correlation and its statistical significance

We know that weight is related to height; that taller people tend to weigh more than shorter people. There are exceptions, but we are talking about a general rule. It is not a mathematical relationship; if someone tells us their height we cannot calculate their weight by applying a formula, but there is a tendency, a specific relationship.

The following graph shows the relation between the weight and height of a group of 92 university students (the data was obtained from a file included in statistical software package Minitab, the same one mentioned in Chapter 1).

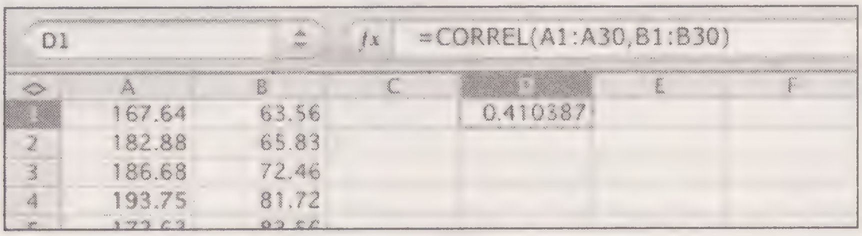


The relationship between weight and height of a group of 92 students.

Would you say there is a 'strong', 'medium' or 'any' relation? As you will see, we need to be more specific in order to evaluate these types of situations, to do this there is a measurement called 'coefficient of correlation'.

The formula for the coefficient of correlation is a little fancy but easy to justify (don't worry, we won't go into it here). In regards to other possible alternatives that exist, the coefficient of correlation has many advantages. Its values are always between -1 and 1, and it does not depend on the units in which the data was collected. In our case, it gives the same result if we have them in centimetres and kilos as if we had them in inches and pounds (which is how they are in the original file).

If the coefficient of correlation is 1 it means that the relation between the two variables is perfect, and that when one is increased, the other increases too. In this case we do have a mathematical relation and the exact value of one variable can be calculated from the other, but with real data such perfection should never be expected. If, for example, it is 0.8 it generally means that there is a clear relation; we get 0.785 for our data. If it is zero it means that there is no relation. The negative values indicate the same as the positive ones, but in this case when one value increases the other decreases.



D1 =CORREL(A1:A30,B1:B30)

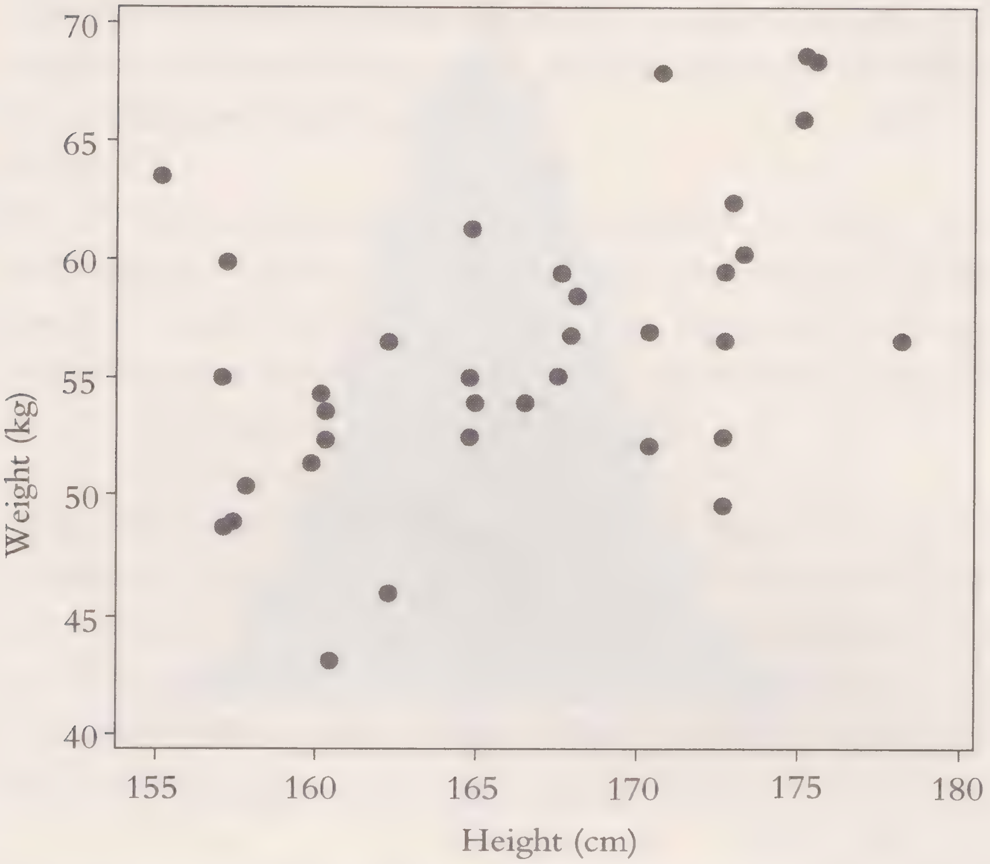
	A	B	C	D	E	F
1	167.64	63.56		0.410387		
2	182.88	65.83				
3	186.68	72.46				
4	193.75	81.72				

The Excel calculation of the coefficient of correlation.

But this measurement also throws up a few problems (nothing is perfect!). If there is no relation between the variables we should expect the coefficient of correlation to have a value of exactly zero, but this means that the data is distributed with a perfect balance, which never happens in practice. What we can say is that it will be around zero. Now the problem is: what does "around zero" mean?

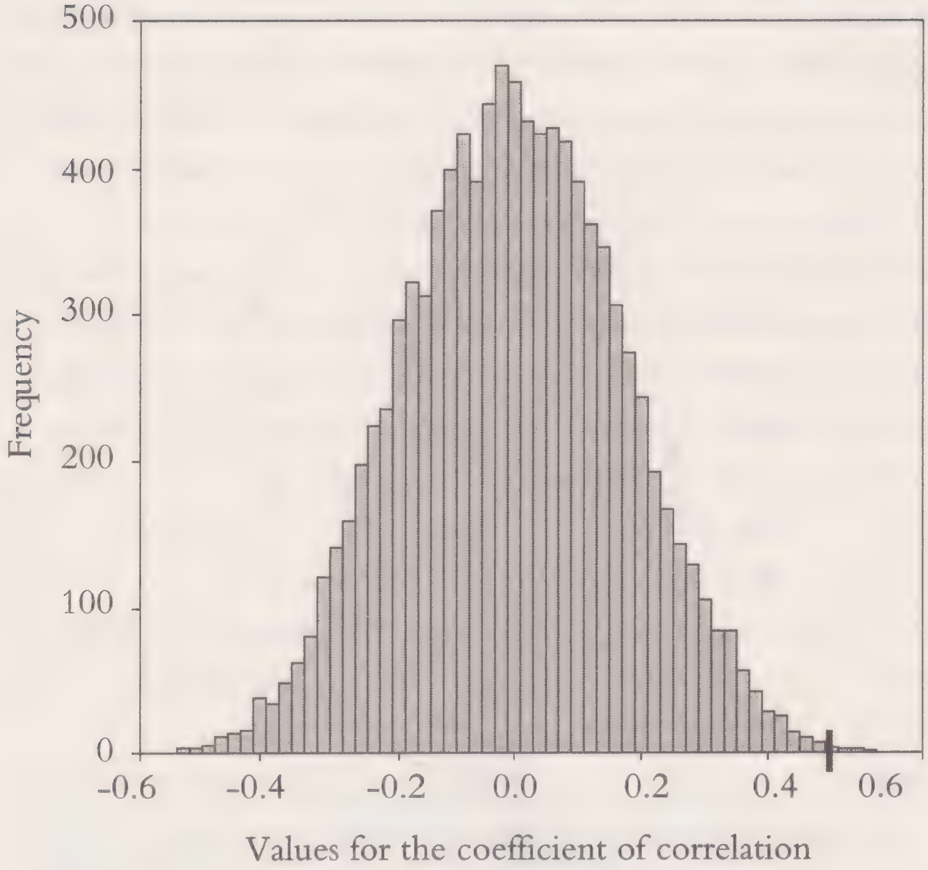
It also has the complication that its meaning depends on the number of points we have. If we have very few points and the coefficient is a long way from zero it is a very unreliable indication of the possible existence of correlation. If there are only 2 points, we simply get 1 or -1, whether there is a correlation or not.

The following diagram contains 35 points, and the coefficient of correlation is 0.494. Is this far enough from zero to be able to confirm the existence of the correlation? Or is it more logical to think that this distribution of points (or the coefficient of correlation, which is the same thing) can be obtained at random without there being any kind of relation between the variables.



Is there a relationship between these variables?

In order to see if it is reasonable to take the coefficient of correlation obtained seriously (technically we would say 'to see if it is statistically significant') we can draw on simulation. We generate two sets of random numbers, 35 on one side and 35 on the other. It is clear that these two sets of numbers will have no relation between them; as they were chosen at random they will be completely independent, but we know that their coefficient of correlation will not be exactly equal to zero. It will be, for example, -0.123, and if we repeat the process generating two sets of 35 random numbers again the new coefficient of correlation might be 0.213, and if we do it again... And if we do it 10,000 times we will get 10,000 values for the coefficient of correlation for two sets of 35 values which are completely unrelated. Doing this and noting down the results would be very time consuming, but there is a little program that does it almost instantly. The results are represented in the following histogram, in which, the value under discussion is indicated by a line.



Values of the coefficient of correlation for sets of 35 pairs of independent data.

We can see that our value can come up when the variables are independent, but also that it is very unusual for this to happen. Analysing the results of the simulation (which cannot be seen in the histogram) we have 12 values above 0.494 and 9 below 0.494. This means that a variation from zero such as the one we have here, or one greater than it, comes up approximately 2 times in every 1,000 when the variables are independent of each other. Is our case one of those 2 in every 1,000? We do not know, but it is not very likely. The most logical thing is to consider that there is a relation between our data sets which, by the way, correspond to the weight and height of 35 of the women in the group of 92 students who we considered previously.

Line of reasoning: contrasting hypotheses

Both in the case of the tea taster and the case of the relation between the variables we should ask ourselves: is it reasonable to believe that the tea taster can tell the difference between one mixture and the other? Can we consider that the two variables are correlated? In both cases the reasoning followed should be:

1. We come up with a default hypothesis, which tends to be conservative. In the case of the tea taster, we assume that she is not capable of telling the difference between the two types of preparation, and in the case of the correlation, we start by assuming that there is none.

2. We calculate a value from the available data, and if we have no data (or that which we have is of no use) we have to obtain some. In the case of the relation between variables, the value that summarises the situation is the coefficient of correlation. In the case of the tea taster, a test needs to be planned and the result is the number of errors committed.

3. If the value obtained is among those that are expected when the default hypothesis is correct, there is no reason to say that it is incorrect and, therefore, we keep it. However, if the value is not very probable, if it does not match the default hypothesis, we are left with the alternative (the lady can tell the difference between the two types of tea, there is a relation between the variables).

In statistics texts you will see that the default hypothesis is called the 'null hypothesis' and the alternative (in other words if the default hypothesis is not credible) is called the 'alternative hypothesis' (no surprises there). The probability of getting a value such as the one obtained (or an even more dissenting one) if the null hypothesis is correct, is called 'p-value', and this is the number that is most used in statistical tests, because it holds the key to whether it is reasonable to maintain or reject the null hypothesis.

In our cases, if the tea taster correctly identifies the 4 mugs of one type we will be able to reject the null hypothesis with a p-value of 1.4%. In the case of the relation between variables the p-value is 2%, as if there were no relation between them (null hypothesis) the probability of having a coefficient of correlation such as the one we obtained, or greater, is exactly 2%.

What if the null hypothesis cannot be rejected?

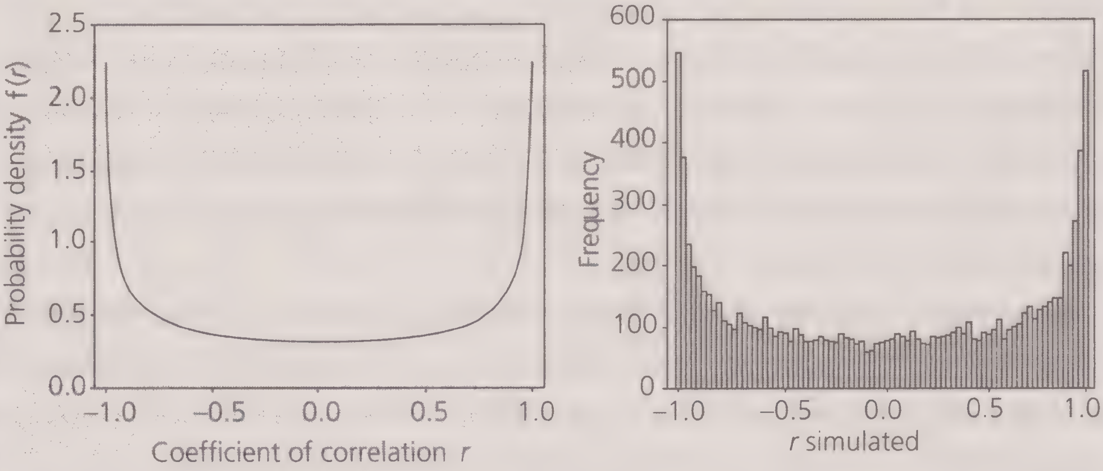
If the p-value is large we cannot say that the data is contrary to the null hypothesis, but that does not mean, in any way, that it has been demonstrated to be true. This is why we prefer to talk of rejecting, or not, the null hypothesis and not of accepting it (perhaps this is a level of subtlety that is often not understood) and, we certainly do not talk of having demonstrated that it is correct.

A simile that is always used to explain the situation is that of court cases, in which, as we know, the null hypothesis is that the defendant is innocent. In other words, they are considered innocent as long as there is no evidence to suggest

AN UNUSUAL CASE: DISTRIBUTION OF THE COEFFICIENT OF CORRELATION WITH 3 POINTS

Fisher was the first person to find a general formula for the distribution of the coefficient of correlation. The mathematics he used are not at all basic, and it seems as if Karl Pearson, another of the greats of statistics and editor of the leading magazine of his time, did not understand it and he discussed its pitfalls in his publication. Fisher did not like this at all, and the incident lead to animosity and rivalry between undoubtedly the greatest statisticians of their time (which, on the other hand, is probably not so strange).

The formula provides strange results. If we have 3 points corresponding to independent variables, the distribution of the values that their coefficient of correlation can take has a strange shape, exactly the opposite of the omnipresent bell shape: the more probable values are those along both edges.



Theoretical distribution of the coefficient of correlation of independent variables with only 3 points, according to the theoretical formula deduced by Fisher (left) and resulting from 10,000 simulations (right).

If there are 4 points then all the values of the coefficient of correlation are equally probable. For 5 points the most frequent value is now zero and as the number of points increases the 'inevitable' bell shape begins to appear.

otherwise. The evidence that has been collected is proof that may or may not be contrary to the hypothesis of innocence. If the victim's blood were found on the defendant's clothes there is evidence contrary to the hypothesis of innocence, but if there is not, if there is no evidence because the crime was very well planned, or because the police did a bad job, the defendant cannot be punished (the null hypothesis cannot be rejected), but that does not mean that innocence has been demonstrated.

Another example: were the dice equally balanced?

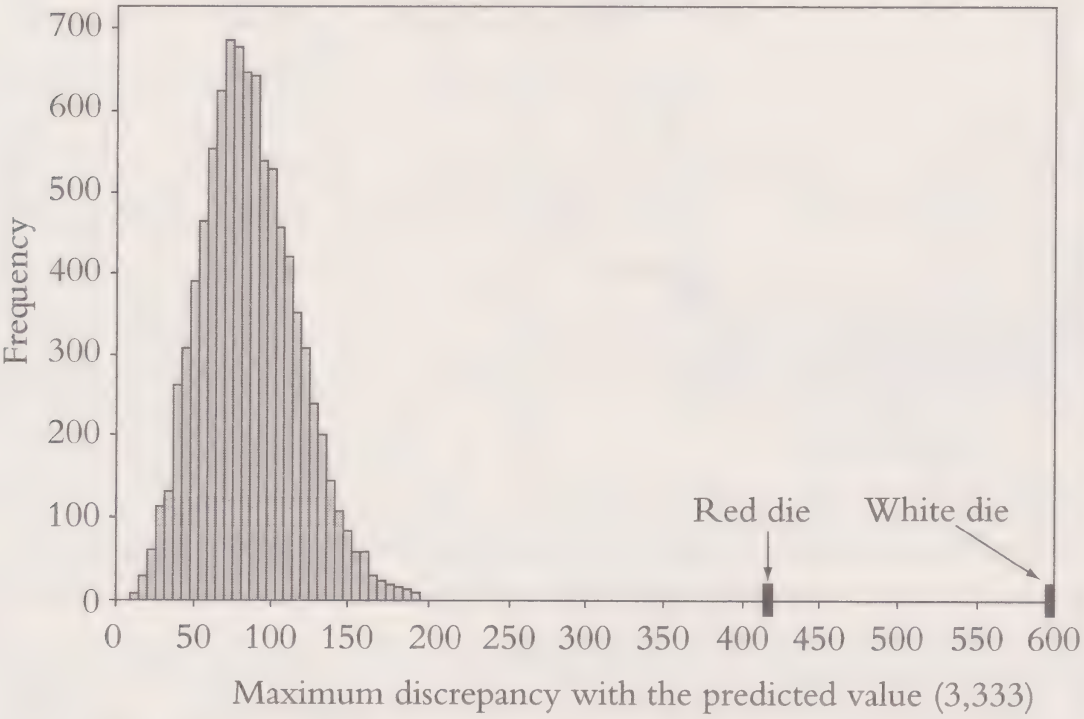
In Chapter 2 we mentioned that in 1850 a Swiss astronomer amused himself by throwing a couple of dice 20,000 times, one red one and one white one, and that in both cases the results obtained seemed to be quite different to the expected theoretical results, which made us suspect that the dice were not equally balanced. As each of the possible 6 results is equally probable, if he threw the dice 20,000 times, the expected theoretical result for each of the 6 possible results is 3,333 (20,000/6). The following table shows the theoretical and obtained values and the absolute value of the discrepancy.

		Results					
		1	2	3	4	5	6
Red die	Obtained value	3,407	3,631	3,176	2,916	3,448	3,422
	Theoretical value (balanced die)	3,333	3,333	3,333	3,333	3,333	3,333
	Discrepancy (absolute value)	74	298	157	417	115	89
White die	Obtained value	3,246	3,449	2,897	2,841	3,635	3,932
	Theoretical value (balanced die)	3,333	3,333	3,333	3,333	3,333	3,333
	Discrepancy (absolute value)	87	116	436	492	302	599

Are these discrepancies sufficient reason to suspect that the dice are not balanced? Or can it be attributed to chance? Of course it would also be strange if every result were to appear exactly one sixth of the time. To clear up any doubts we are going to carry out a contrasting hypothesis following the reasoning plan that Fisher used in the case of the tea taster. We will start by assuming that the dice are balanced (what else could we assume?) and we will only reject this option if the data we have is contrary to it.

We will take the maximum discrepancy between the obtained and expected values as a relevant value to summarise the information available. In the table below we can see that for the red die it is 417 and for the white one, 599. The question now is: Which values should we expect for these discrepancies if the die is perfectly balanced? Again, we can answer this question using simulations.

We will simulate the throwing of 20,000 dice, count how many times each value comes up and we will get the value which shows the greatest discrepancy from the expected value. The first time we did this the maximum discrepancy was 83, the second time 97 and after doing it 10,000 times, the histogram of the values obtained is the diagram shown below, and we have added a marker for the values corresponding to the red and white dice.



The distribution of the maximum discrepancy when the die is balanced and the values that were actually obtained.

It is clear that our data is contrary to the hypothesis that the dice are balanced. If it were true, values like those we have would by very very unlikely.

The p-value would undoubtedly be zero with many decimal places and, therefore, with the probability of being wrong practically null, we can confirm that the dice were not balanced.

Instead of only choosing the maximum discrepancy as the value that summarises the information available, we can also use a measurement that takes into account the discrepancies of all 6 results.

This measurement could be the sum of all discrepancies calculated as the difference between the observed and expected frequencies, elevated to their square (so that the positive and negative values do not cancel each other out) and divided by the expected frequency.

In other words, for the red die:

$$\frac{(3{,}407-3{,}333.33)^2}{3{,}333.33}+\frac{(3{,}631-3{,}333.33)^2}{3{,}333.33}+\frac{(3{,}176-3{,}333.33)^2}{3{,}333.33}+$$
$$+\frac{(2{,}196-3{,}333.33)^2}{3{,}333.33}+\frac{(3{,}448-3{,}333.33)^2}{3{,}333.33}+\frac{(3{,}422-3{,}333.33)^2}{3{,}333.33}=94.189.$$

This measurement may seem unnecessarily complicated, but it has the advantage that it is not necessary to construct the distribution by simulation which follows when the null hypothesis is correct (we call it the 'reference distribution'). The distribution that follows this measure of discrepancy is very well known and has a name that is rarely forgotten by those who have heard of it. It is called 'chi-squared' and this type of test is called a 'chi-squared test'. It was first used in 1900 by Karl Pearson, another of important characters in the history of statistics (His name is sometimes used to describe the coefficient, the full term being 'Pearson's coefficient of correlation').

For the most common statistical tests it is not necessary to obtain the reference distribution by simulation, instead it is deduced with mathematical reasoning. The formula that gives the distribution of the coefficient of correlation is quite complicated and has no specific name, although if the sample is large it looks a lot like a normal distribution. By the way, the first person to deduce the formula for this distribution was… Ronald Aylmer Fisher.

LITTLE DISCREPANCY IS ALSO SUSPECT

If we throw a perfectly balanced die 20,000 times, each of the 6 possible results will appear around 20,000/6 = 3,333 times. It is very rare that the discrepancy between the observed and theoretical frequencies are more than 250 for any result. This only occurs in around once in 100,000 times.

But it is also very unusual for the frequencies obtained to look very like those predicted. For example, if someone told us that they had obtained the following results from throwing a die 20,000 times:

1	2	3	4	5	6
3,333	3,334	3,333	3,333	3,334	3,333

we would have reason to doubt the authenticity of that information, as such a similarity between the expected and obtained frequencies occurs less than one in a million times.

Fisher demonstrated an interesting coincidence between the experimental data published by Mendel in his famous work on the genetics of pea plants and the theoretical results that would have been expected. The most surprising thing is that Mendel had predicted an incorrect result for some experiments and the experimental results showed a suspicious similarity with those incorrect values. It was not necessarily Mendel himself who falsified the data, Fisher said, but an assistant who had not done their work properly and knew what Mendel wanted to hear...

This has been the subject of some in-depth discussions. It is not just a problem of calculating probabilities but also of genetics and botany when discussing the possible incidences that could arise in plants and what leads to variations in the proportions obtained in other types. The controversy has gone on for a long time and it seems difficult to find definitive conclusions, although there is a general consensus that there is no solid evidence that Mendel, or whoever it was, fixed the data.

Until now, yes; from now on, no: boundaries for the p-value

A value is normally set, often 5%, such that if the p-value obtained is lower, the null hypothesis is rejected, and if it is greater, it is not. This boundary value is called the 'significance level'.

Although we all like clear and simple rules, it is not easy to take a value as a universal boundary and always apply it regardless of the context of the situation. Setting a boundary value is the same as deciding the probability of being wrong if the null hypothesis is rejected, and the probability of error that can be reasonably assumed depends, undoubtedly, on the situation we find ourselves in and the consequences of committing an error.

Let's assume, for example, that one morning we check the weather when leaving our house and find that there is a 10% chance of rain. Should we go back home to get an umbrella? Nobody would consider it reckless if we carried on and ran the 10% risk of being caught in a shower. If we are wrong we do not lose much (perhaps we would get a little wet) and we should also consider that it is quite annoying to carry an umbrella around all day if it does not rain.

Another situation. We are driving on a quiet main road. At the brow of a hill, with no visibility of oncoming cars, we see that there is a small dip in the place where we have to drive, although we could avoid it by driving on the other side of the road. But we do not do it. The probability of finding another car on that little used road is small, and finding it just at the brow of the hill is even less likely. But we do not do it because, although the probability is small, if it does happen, the result would be very harmful indeed. Also, passing over the dip is only a slight inconvenience.

It is evident that the probability of error that we are willing to risk when taking a decision depends on the circumstances and on the cost of being wrong.

Staying on the road for our next example, but this time a less dramatic one, we could consider the case of speed guns for measuring the speed of cars. It is well known that these devices, like any other, give results with a certain error of measurement, so that if they show that a car is going at 70mph it is also possible that it is going at 69 or 72. This is why if the maximum velocity is 70mph, drivers are only fined if the radar shows a speed which exceeds that limit by a certain amount, so that despite the inevitable error in measurement it is practically certain that the driver is breaking the limit. Choosing a margin above the limit that gives 5% errors (fines for drivers who have not actually broken the speed limit) would be a careless decision, as it means that every day hundreds of drivers would be unfairly fined.

In short, the selection of the boundary value is not a statistical problem, instead it depends on the problem we are dealing with. When a test is carried out to analyse if a new drug is better than the current one for curing an illness, taking 0.05 as the border value means that we run a 5% risk of saying it is more efficient when it actually is not. What are the implications of this? Could the new treatment have harmful side effects? Is it much more expensive than the conventional treatment? The answer to these questions are important for setting the most appropriate boundary value.

But it is also true that in many cases the value of 0.05 is taken as a reference without going into detailed reasoning. The reason for using 0.05 is to do with the values that appear in the tables. When they were first created, with some very rudimentary methods, only the values corresponding to a few probabilities were included – easy numbers such as 0.001, 0.005, 0.01, 0.05, 0.10… – and from those available it was customary to take the value corresponding to 0.05 as the most suitable for separating the normal from the unusual. The advantage of 0.05 is that it is a rounded value in our decimal system. If we had 6 fingers we would undoubtedly consider it more natural to make decisions using 0.06 as the boundary.

Chapter 5

Better? More Effective? How to Design Samples to Answer Questions

Statistics is necessary when a question is asked and we need to gather and analyse data to find the answer. Questions of this type include those that ask whether a vaccine is effective, whether a medicine is better or if one welding system is stronger than another.

One of the common problems is that the process of obtaining data is always laborious (and expensive), and it is necessary to think of the best way to do this in order to avoid wasting what are always scarce resources. Another problem is that in the end we never have all the data we would like to have and it is necessary to take full advantage it. All this against the backdrop of variability: the data is not determined by a mathematical expression and under the same conditions, we do not always get the same result.

Imagine the following question: is it helpful to take a certain dose of aspirin regularly to reduce the possibility of having a heart attack? We can try to answer by reasoning about the effects of aspirin on the body, but the truth is often surprising. The most reliable solution is to gather data. Essentially, this involves dividing a group into two parts that are as similar as possible: One part is assigned the aspirin treatment and the other is not. The results are then compared. We know that not all of the individuals who take part in the study are the same; we also know that not all will react in the same way when they take aspirin. It is necessary to know how to deal with all this and reach conclusions that also indicate the degree of reliability which has been achieved. This is statistics.

A large-scale study: the polio vaccine

The possibility of being vaccinated against an infectious disease has without doubt been one of the discoveries that has had the most impact in the fight against illnesses and in improving health and life expectancy. However, each illness requires its own specific vaccine and finding this is not always an easy task. There are various procedures for preparing them and laboratory experiments or animal or human testing on a small-scale can often provide quite a few clues as to their level of effectiveness. However, prior to approving a vaccine and recommending large-scale use on an entire population, we must be extremely sure that the benefits compensate for the costs and risks that will inevitably be assumed. And statistics has a lot to say when it comes to such verification.

In 1954 a large-scale study was carried out to evaluate the effectiveness of a vaccine against polio (the Salk vaccine developed by the epidemiologist Jonas Salk). The process followed is clearly explained in the book *Statistics: A Guide to the Unknown*, which provides 29 case studies of applied statistics in a wide range of areas, each written by an author with in-depth knowledge of the field in question. The chapter on the work carried out to verify the effectiveness of this vaccine was written by professor Paul Meier from the University of Chicago.

The interest and peculiarities of polio

The effectiveness of the vaccines that have been developed have almost resulted in the complete eradication of polio, which until recently was one of the most feared illnesses. It attacked children, leaving many paralysed or with long-term damage for the rest of their lives. In addition to this, it came in waves of unpredictable epidemics and, rather curiously, affected social groups with higher standards of living, while hardly occurring among the poorest countries or stratas of society. This was because the attack occurred earlier among poorer populations, when babies were still protected by the defences of their mother, meaning that when they were attacked by the virus, they did not develop the illness and were thus also immunised. On the other hand, those who lived in the most favourable conditions contracted the disease later, when they could no longer count on their mother's defences. Another circumstance that surely had an influence on the fight against the disease was that President Roosevelt had also suffered and was extremely willing to support research in the field.

At the start of the 1950s, the United States health authorities believed the safety and effectiveness of the new vaccine developed by Jonas Salk had been proven in small-scale studies. However before recommending large-scale use it was necessary to have irrefutable evidence both that it was effective and that it did not have damaging side effects. This formed the context of the largest experiment that had been carried out in the field of public health at that time.

The control group

Let us assume that a medicine for a certain illness is trialled and it is confirmed that anyone who takes it will be cured within 7 days. Can we then say it is effective?

Perhaps you will be thinking yes: if everyone is cured, it must be effective. However, the truth is that this type of experiment does not show this. It is possible that without taking anything, those people would also have recovered during the 7 day period. It could even be the case that without taking the medicine, they are cured in 2 or 3 days, while taking it means waiting 5 or 6.

For this reason, when the effectiveness of a new medicine or vaccine is tested, it is necessary to start with a set of individuals who are representative of those against which the vaccine is directed and divide them into two random groups to ensure that there are no systematic differences between the characteristics of those in one group or another. The medicine will only be given to those in one group; those in the other will be used for a comparative analysis of the effects of the new drug. This group will not receive any treatment and will be referred to as the 'control group'.

In the case of polio, its incidence rate displayed unpredictable fluctuations. For example, in 1952, the year with the greatest incidence between 1930 and 1956, there were around 60,000 people affected in the United States, whereas in 1953 there were only 35,000, a reduction of more than 40%. If in 1953 a new and completely ineffective vaccine had been trialled, it could have been believed that it was effective in light of the considerable reduction in cases. And this was not an exceptional case, from 1931 to 1932 the number of cases dropped by more than half, and the same thing happened from '35 to '36, '37 to '38, '41 to '42, '46 to '47 and '55 to '56.

Nor was it a good idea to vaccinate children from a specific area – vaccinating those from the state of New York, for example, and leaving those from another area, such as Chicago – since the incidence rate was not homogeneous and it could easily be the case that in one year, the incidence in one state was high whereas in another it was low. It was necessary to divide all the participants in the study into two similar groups that would be affected by all these factors in the same way. One group was supplied the vaccine and the other acted as the control.

Two groups 'as similar as possible': control with placebo and 'double blind'

If one group of people receive a treatment (they take a pill every day or receive a single injection, as was the case with the Salk vaccine) and another does not receive anything, those who have received something, if they are convinced that it has a curative effect, will notice some form of improvement, even when the product is completely lacking in therapeutic action. This is referred to as the 'placebo effect'. Surely herein lies the success of many so-called alternative medicines and the fact that many problems will go away of their own accord, with or without treatment.

In the case of polio, the child is either affected by the illness or is not, and it could be thought that as such there is no problem regarding the improvement perceived depending on whether the participant has been vaccinated. However, not all cases are serious or have long-term effects, and if a child who has been vaccinated shows symptoms that may be caused by polio, perhaps the parents and also the doctor, may think that the child has been affected (even though they were vaccinated!). If it really was a mild case which developed it could be confused with another illness and end up being an unrecorded case of infection. On the other hand, those who have not received the vaccine and who are more attentive to any symptoms as a result of feeling unprotected, will surely analyse the symptoms more carefully and will be diagnosed, with the possibility of a false impression that there was a higher incidence rate in the group that did not receive the vaccine.

In order to avoid the placebo affect only working in favour of the group that is treated, in this type of test it is best for all participants to receive a treatment with a similar appearance so that they will be unaware if they are taking the active ingredient or the false one referred to as the 'placebo', which has the same appearance and taste as the real pill. However, not only is the individual participating in the study unaware of whether they have been included in the treatment or control group (in the case of polio, instead of the child, perhaps we should think of its parents), but the doctor who is treating them is also unaware of whether they are taking the placebo or the active ingredient. It is not that we cannot trust doctors, but it could be that they are lead by their prejudgements, and if the child is part of a group receiving the treatment and believes this to be effective, they will tend to report greater improvement, whereas if they know that the child has taken the placebo, perhaps they will tend to interpret what the patient tells them with a greater emphasis on the negative.

SIGNIFICANT AND IMPORTANT DIFFERENCE

When we make comparisons, we attach great importance to deciding whether the differences which are observed are significant or not. In fact, this is what all statistical tests set out to clarify. However, although it may seem like a contradiction in terms, the fact that a difference is significant does not necessarily imply that it is important.

A difference is said to be significant when it is believed that it cannot have been caused as a result of chance, or in other words, that the two treatments being compared really do give different results. Nonetheless we can be sure that they are different, although this difference may be so small as to be irrelevant for practical purposes.

For example, a study carried out with many samples may show that a certain type of glue is stronger, but that the difference is almost imperceptible. It may also be the case that as a result of having little data or due to the high variability of the results, a large difference is observed but that this can be attributed to chance. In short, we are unsure whether one is better than the other.

To ensure there is no possibility of this happening, this type of study is designed in such a way that neither the patient nor the doctor treating them and reporting the results know who is taking the active ingredient and who is taking the placebo. This is why they are referred to as 'double blind'.

However with a control group taking a placebo, certain problems also arise, one of which is that the experiment is more complex to organise. In the case of the Salk vaccine, it was necessary to prepare injections with the active ingredient that were identical to those that only contained a saline solution, meaning that they had to be numbered and controlled in order to identify which was which, in spite of the fact that not even the health workers administering the vaccine could know if it contained the active ingredient.

Another problem lies in the realm of ethics. To some people it did not seem fair to inject children who were participating in the study with a saline solution instead of a vaccine that people were already extremely sure was effective.

As an alternative, it was suggested that second-year children should be vaccinated and the first and third year ones used as the control group. This resulted in certain problems, such as breaking the double blind principle, but was eventually used in approximately half of the areas, whereas the other half used placebo control groups.

The requirement for an extremely large sample

The incidence rate of the illness was only 50 per 100,000 individuals, and it was hoped that the vaccine would reduce this number by half. It is clear that the trials cannot be carried out using small groups. For example, if we vaccinate 1,000 children and another 1,000 are used as a control group, it is most likely that there will not be anyone affected in either of the groups and the experiment will not have been of any use. If 10,000 are used, it may be the case that in the control group there are 5 people affected and in the vaccinated one, only 2. But the difference is so small that it can be attributed to chance. (It will not be possible to reject the null hypothesis that the rate of effectiveness is the same.) It was necessary to have hundreds of thousands in each group before the conclusions could be considered to be solid. It was necessary to carry out a large-scale experiment.

Results

It was verified that the vaccine was unquestionably effective. In the group that had received the vaccination, the incidence rate was less than half that of the group that had been injected with the placebo. The p-value of this trial was of the order 10^{-9}. In the event that there were no differences between one group and another, a discrepancy of this size could only arise by chance was approximately one in a billion.

	Population studied	Cases of Polio	
		No.	Rate (per 100,000)
Control with placebo:			
Vaccinated	200,745	57	28
Placebo	201,229	142	71
Control by school year:			
Vaccinated	221,998	56	25
Control groups	725,173	391	54

The results were similar in the areas where the preceding and following school years were used as a control group and everyone was satisfied with the way in which the experiment had been carried out, a good result all round. However, in spite of the fact that the difference was clearly in favour of the group that had received the treatment, without the placebo control group, there would always have been doubt as to whether other interpretations of the results were possible.

The role of statistics: polio today

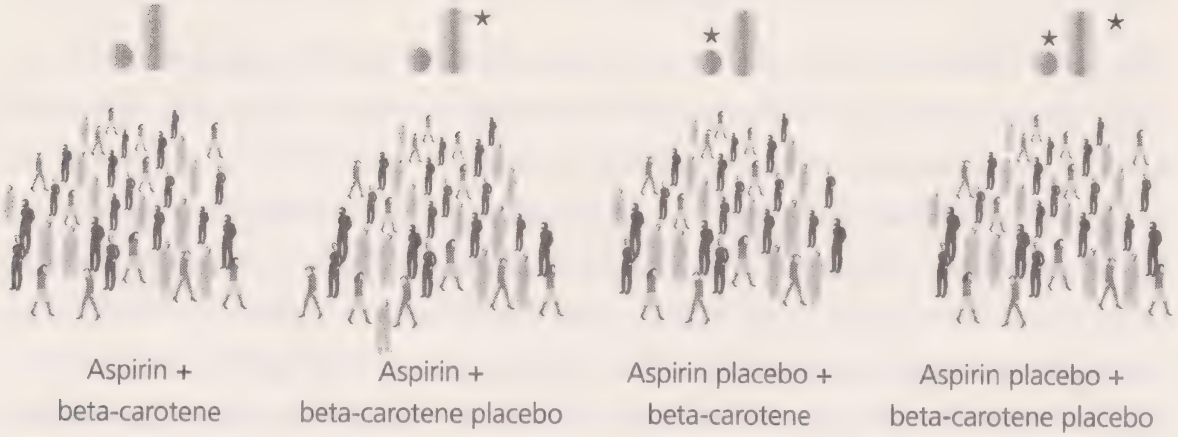
The Salk vaccine, while representing a step forward in the fight against the disease, was not wholly satisfactory and a few years later was substituted for another, more effective vaccine, which underwent statistical trials that were appropriately designed and executed prior to being used. Today, Polio is a disease on the way to becoming extinct. There are only four countries in the world in which it continues to be endemic: Nigeria, India, Pakistan and Afghanistan. The WHO, UNICEF and other international organisations have announced that they are making efforts in these countries and estimate that new cases of the illness will soon cease to occur. It will then be necessary to wait for 3 years until polio is officially declared to have been wiped from the face of the earth.

Aspirin and heart attacks

In 1983 a large-scale study was undertaken in the United States to measure the influence of aspirin on heart disease. Smaller studies have shown that a person who had suffered from a heart attack could reduce the risk of a second one by taking aspirin. However, there was no proof that this beneficial effect could be extended to the male population in general.

261,248 male doctors over the age of 40 were invited to take part, their details having been obtained from the American Association of Doctors. From these, 59,285 offered to participate in the study although it was necessary to exclude candidates with a complicated clinical history, who were already taking aspirin or reacted badly when they took it. This left a total of 22,071 doctors who could all be considered healthy, low risk individuals, and they were administered a dose of 325mg of aspirin (or the placebo) on alternate days.

While studying the effect of aspirin, the scientists behind the study took advantage of the opportunity to also study the effect of beta-carotene (a compound which the body transforms into vitamin A) on preventing certain types of cancer. Thus the participants were randomly arranged into four groups which took real aspirin and beta-carotene, real aspirin and a beta-carotene placebo, a placebo aspirin and real beta-carotene, and placebos for both aspirin and beta-carotene.


Aspirin + beta-carotene
Aspirin + beta-carotene placebo
Aspirin placebo + beta-carotene
Aspirin placebo + beta-carotene placebo

Treatments assigned to the four groups. The appearance of what they took was identical in all cases. The asterisk indicates what was really just a placebo.

In spite of the strict selection criteria applied to participation in the study, it involved a set of people of various ages, different clinical histories, different characters, some smokers and others not… and as such it was necessary to be extremely scrupulous in the random allocation of each individual to one of the four groups, since only in this way would it be possible to ensure that the four had similar characteristics. It could be suggested that the participants who were on the verge of suffering a heart attack may have been clustered in one group, but probability theory assures us that if the distribution is genuinely random, the probability of this happening is so large as to be completely negligible.

Thus, as the four groups were made up of similar subjects and were subject to similar environmental conditions, if important differences were observed in the results of each group over and above what could be attributed to chance, they could be attributed to the fact that the treatments that had been followed had produced different results. This is the logic of experiments which compare treatments with independent random samples.

The study was double blind, meaning that neither the participants nor those responsible for monitoring them knew to which of the four groups they had been assigned. A supervisory committee analysed the results every six months and although the study had been planned to run for seven years, after just five, the results were so good that the research was stopped so as to report back as quickly as possible to the participants and the medical profession in general.

	Aspirin group (n = 11,037)	Placebo groups (n = 11,034)
Heart attacks		
Fatal	5	18
Non-fatal	99	171
Total	104	171

The aspirin group was made up of people who took aspirin and beta-carotene and those who took aspirin and the beta-carotene placebo, and the placebo group of the other two. A statistical test which compares proportions clearly shows that if the aspirin had no effect (the probability of having a heart attack was the same in both groups), a difference like the one we have obtained, or higher, would only occur at random in the order of twice every 100,000. It is reasonable then, to consider that aspirin reduces the potential of suffering from a heart attack.

The headline made the front page of the *New York Times* and was widely reported in the media. With respect to the beta-carotene, the study continued for the planned period although I have not been able to find references regarding the final results. I suspect that they were not good – consulting on what is currently known about its effects – it appears that it does not only reduce the possibility of contracting cancer, but increases it in the case of smokers.

But neither is aspirin a panacea. It is believed to act by inhibiting the process for binding platelets and thus reducing the formation of blood clots. However at the same time, this entails a potential danger. In fact the study showed that there was a small but not significant increase in mortality from embolisms among those who received aspirin. This is why the instruction has been taken with a pinch of salt and it is insisted that patients follow the recommendations of their doctors who should evaluate the advantages and drawbacks in each case.

Tobacco and lung cancer

Nowadays there is no example of clearer and more widely accepted evidence than the fact that 'smoking seriously damages your health'. However this was not always the case. We now know that components of tobacco smoke can cause cancer and we know how they act to induce the transformation of healthy cells into cancerous ones, something which has been proven in experiments using animals. However, just like on so many other occasions, statistics sounded the alarm that something was not right and this led to more detailed research. The data available in the 1950s already showed a higher incidence of lung cancer among smokers than among non-smokers, but it was necessary to conduct more detailed studies to confirm these suspicions.

To verify the connection between lung cancer and other illnesses with smoking, seven large-scale studies were carried out in English-speaking countries (one in the UK, another in Canada, and five in the United States). All were large-scale studies in which the number of participants varied between 34,000 and 448,000. The procedure was essentially the same in all the experiments: a questionnaire was sent to the people selected to participate, asking them about current and past patterns of consumption alongside other demographic data, and a system was created to ensure that when one of the people who had responded to the questionnaire died, the fact would be recorded together with the cause of death.

These studies made it possible to have data on the influence of factors such as the age at which participants take up smoking, the type and quantity of tobacco they normally smoke, as well as what happens to ex-smokers. One of the conclusions was that among smokers, the incidence of lung cancer is between 11 and 20 times greater than among non-smokers.

One of the objections that can be raised to this type of study and which was in fact made (and Fisher was one of the champions), is that while they show that there is a higher incidence of, say lung cancer, among smokers, this does not imply that the cause is tobacco. For instance it could be the case that those who acquire the habit of smoking are more restless, more nervous and that this same trait which makes them liable to take up smoking also makes them liable to suffer from such illnesses, or it could also be that there is something in the genetic structure of certain people that makes them susceptible to becoming addicted to tobacco and, at the same time, but not as a consequence, gives them a higher probability of suffering from lung cancer.

These objections can be raised because the studies which were carried out were not experimental designs like in the case of the polio vaccine or the studies for the effects of aspirin on heart disease. In the latter cases, the study group was randomly divided into two parts, the treatment group and the control group, and as everything was designed so that the only difference between one group and another was the factor under study, if there were significant differences between the groups, these would necessarily be caused by the factor that differed between the groups. However the studies on the influence of tobacco were not experimental designs but prospective studies, or rather observing how two previously existing groups evolved with respect to each other. In this case, it was not possible to force non-smokers to smoke and to stop compulsive smokers from smoking. Taking the theory to an extreme, it would have been ideal for all the participants to smoke, with a randomly selected half smoking normal tobacco and the other half a substance which was known to be completely innocuous and which had an appearance and 'taste' identical to tobacco.

Tobacco manufacturers could argue that this would be the only correct study, and they would be right, but such an experiment is as correct as it is impossible. But from the available data, it is clear beyond any reasonable doubt, that tobacco is a significant risk factor in terms of lung cancer, bladder cancer and heart disease, as well as other types of illnesses. The connection between lung cancer and the consumption of tobacco has been observed in multiple studies in diverse countries and contexts, thus eliminating any possible predisposition of a specific group of people, but furthermore, we now know which components of tobacco smoke can cause cancer. The genetic hypothesis cannot explain the increase in the incidence rate among women when they have taken up smoking, or the increase among non-smokers exposed to second-hand cigarette smoke. In short, it is obvious, although this was not always the case, and statistics has been at the cutting edge of providing arguments to make this clear.

Randomisation and blocking

When experiments are designed to compare what are generically described as *treatments* (these could be to compare two medicines against an illness or two catalysts to improve the performance of a chemical reaction) the key lies in being able to obtain two sets of data in which the only variable that affects the outcome is that which is being studied. In the field of medicine, we can compare two medicines, or taking a medicine as opposed to not taking anything, as was the case with the polio vaccine or the influence of taking aspirin on heart attacks. As we have already seen, the key lies in dividing the participants in the study into two groups which are as similar as possible (chance, although it may seem paradoxical, is a good way of ensuring this balance between the two groups) and which are affected in the same way by all factors except the one whose effect we wish to study. As such, if there are significant differences between the two groups (differences beyond what can be reasonably attributed to chance) these are attributed to the factor which has acted differently on the two groups. However, if in addition to the factor being studied, others affect each group differently, in the event that there are differences between one group and the other, it is impossible to know if these are a result of the factor that is being studied or any of the others that are also acting.

Let us consider an example. One of the standard texts on experiment design is *Statistics for Experimenters* by Box, Hunter and Hunter, which explains how an experiment can be designed to compare the wear of two materials for manufacturing the soles of shoes for young people. If there are, for example, 10 young people in the trial, one idea would be to randomly divide them into groups of 5: one group is given shoes with sole type A and the other group is given shoes with sole type B. After a given period (e.g. 6 months), they are asked to return their shoes. The wear of the soles manufactured from each material is then measured and the corresponding statistical analysis is carried out. (In this case it will be the so-called Student t test for independent samples).

Naturally, the distribution will have been made at random. It's not good enough to go to a school playground and ask the children to line up and give the first five in the line shoes with sole type A and the last five sole type B since the first will probably run more, move more and cause a greater degree of wear to the soles of their shoes.

There is a flaw in this design for gathering data. The wear of the soles of their shoes can be influenced by the material (this is what we are trying to determine), but it can also be influenced by the child: perhaps there are children who run a lot and even play football in these shoes, while others only run a little and only play football on computers. It could even be the case that some of the participants hardly wear the shoes because they don't like them or because they hurt, meaning that they will hardly wear down the soles.

As such, if the wear is not only influenced by the material of the sole, but also by other factors, in the event that a difference is found, it is not possible to know if this is down to the material or the other factors. It could even be the case that no differences are detected as a result of factors which disturb the results, when in fact they do exist.

How can we overcome this problem? By giving each child one shoe with one type of sole and another with the other type. As the two feet always go together, any differences in wear must be attributable to the material and not to any other factor.

This type of design does not compare the average of one sample with respect to the average of another, but instead the comparisons are carried out child by child.

WILLIAM SEALY GOSSET, AKA "STUDENT"

Anyone who has more than a passing interest in the study of statistics will have come across the Student's t distribution (certainly more widely used than the normal distribution) or the Student's t tests for comparing means. "Student" is the pseudonym under which W.S. Gosset (1876-1937), a man who made great contributions to statistics and who developed his career at the Guinness brewery in Dublin, published his work.

At the start of the 20th century, just when Gosset had graduated in mathematics and chemistry from the University of Oxford, Guinness passed into the hands of a young heir who decided to go beyond traditional method and craft and so contracted scientists to introduce more advanced procedures. "Student" was one of them. He quickly realised the importance of using statistical techniques in researching the best ways to brew beer. It was necessary to study the influence of raw materials which could have highly variable properties and which were highly sensitive to environmental conditions. It was necessary to carry out tests, although only a few were ever done and there were always only small quantities of data available.

Until then it was believed that samples were always large enough to be able to provide an accurate estimate of the parameters of the population used to calculate probabilities. However, when using small samples, these estimates lack precision and this changes things. Gosset studied the solution to this problem and published his conclusions in an article under the pseudonym "Student" because company rules forbade its technicians from publishing articles with the results of their research.

There are a number of accounts as to how and why he arrived at this pseudonym. One version states that those at Guinness discovered Gosset's mathematical hobby. However there are also other versions such as one which claims that they were not only aware of these publications but that it was the company's director himself who suggested using Student as a pseudonym. It would seem that their concern was not so much to hide the statistical theories which were being developed but to avoid other brewers knowing that Guinness was using statistical techniques to improve its products and the production processes.

If in general, one sole is worn down more than another (regardless of whether it is worn down a little or a lot since what is important is the difference between one and another), this will be due to the difference between the materials.

The statistical test carried out to compare means when the data has been gathered in this way is referred to as the Student t test for paired samples.

It goes without saying that it would be a bad idea to always put the type A sole on the right foot and the type B on the left one, since perhaps the sole of one foot tends to be worn down more than that of the other, but this can be rectified by randomising the order (e.g. a coin is tossed for each child, if the result is heads, material A is used on the right foot and if the result is tails, it is used on the left one).

In this way, it is expected that even if the foot influences the results, by randomising this influence it is distributed between the two materials without affecting one in a different way from the other.

Randomisation has no cost and guards against the influence of possible known and even unknown factors. For example, a case similar to the one with the soles of the shoe involves studies carried out to check the resistance to deterioration (scratches, wear, etc.) of certain types of coatings which are used on the lenses of glasses. If one group of people is given glasses whose lenses have a certain type of coating and another is given lenses with another type, and after a certain time they are asked to return the glasses to measure the deterioration in the coating, it is clear that the deterioration will be influenced not only by the material but also the treatment each pair of glasses has received as well as the environment in which they are used, among other factors.

As such, just as in the case of the soles, the best solution is to give each participant glasses whose lenses have different coatings (it goes without saying that this is impossible if one is blue and the other yellow). But do we need to randomise the order in which each treatment is applied or can we always apply A to the right eye and B to the left one?

It is always better to randomise. Experts who have studied this issue claim that when we clean our glasses, we always start with the same lens. Not everyone starts with the same eye, but whoever starts with the right will always start with the right and vice versa. And the lens which is always cleaned first is always cleaned more thoroughly… so in short, just in case there was still any doubt, it is always best to randomise.

Try it yourself!

There are urban legends (or perhaps they are not) the truth of which can be tested using statistics. Let us consider some examples.

Is a teaspoon useful for keeping the fizz in a bottle of champagne?

Some people believe that placing a spoon in the mouth of a bottle of sparkling wine prevents the gas from escaping (or at least it does not escape as much as if it were open) and in this way it is better conserved for the following day. One way to clear up any doubts is by... trying it ('trials' are also referred to as 'experiments').

This dilemma recalls the tea taster and can also be tackled using a person who tastes glasses from a bottle which has been conserved using a spoon and another in which nothing has been used. From the outset, we know that the test cannot be carried out using one glass of each type; as a minimum, we need 3 glasses from one bottle and 3 from the other, identical in every way, including the way in which they have been conserved; the only difference will be that a spoon has been placed in one and not the other.

If the taster identifies the 3 glasses that come from the bottle with the spoon, and if everything is done correctly, the probability of them getting it right at random is exactly 5%. (Remember, there are 20 different ways of choosing 3 objects from 6, and only one is correct). For the probability of getting it right at random to be less, we would need to have a tasting with more glasses, but we also need to bear in mind that the taster will doubtlessly gradually lose their sense of taste (as well as others).

We could consider carrying out the experiment with different people, but we would need to be careful because something is always strange if it is tried only once and is no longer so strange on subsequent encounters. If the probability of a person correctly getting it right at random is 5%, if there are 5 people in the experiment, the probability that one gets it right is around 40%, which proves nothing.

It goes without saying that instead of complicating matters in this way, we could just use a machine which measures the content of the gas in the bottles with and without the spoon. Problem solved. It is true, if there were such a machine, this would be a good option, however it is possible that the machine will point to differences and that, for all intents and purposes the teaspoon makes none; after all it is people who drink cava and not machines, and if people are unable to distinguish one from the other, the teaspoon is useless. For the same reason, it is perhaps not such a good idea to hire an exceptionally gifted taster to carry out the tasting.

Do we really know when a melon is ripe?

The selection of a ripe melon is even closer to the case of the tea drinker. There are those who maintain that they are able to choose the best melon based on weight, touch... However, to see if this is really the case, a test could consist of presenting 5 melons chosen at random and asking the expert to choose the best. We would then prepare a tasting with each of the melons (blind, of course) and they would select the best one again, now knowing how each tastes, to allow us to see if the choices are the same. The problem is that the probability of them guessing correctly at random is 1/5 (20%) and as such, even if they are right, this proves nothing. However if the test is carried out twice and they get it right on both occasions, the probability of doing so at random is 4%, and if they get it right 3 times, it drops to 8‰, making it unlikely that they do not have this ability.

Does aspirin make flowers last longer?

It would seem that the virtues of aspirin are not only confined to the realm of medicine. The belief that a bunch of flowers lasts longer if an aspirin is added to the vase is widespread.

An experiment to see if this is true could be organised by taking two bunches of 20 flowers (even better if they are all different) such that we have two roses, two carnations, two daisies... A flower is placed in a vase and its partner in another one, which is the same, positioning them together and ensuring that the lighting conditions and everything that can influence the duration of the flowers has precisely the same effect on both cases. The only difference will be that we will add an aspirin to one jar and not to the other.

If the aspirin has no effect, the probability of one or the other withering first is 50%; as such, it will be highly unlikely that in 20 cases the one with aspirin will last longer. The probability that this happens by chance is the same as tossing a coin 20 times and getting heads, and this is (applying the 'and' rule we saw in chapter 2) $0.5^{20} = 9.5 \cdot 10^{-7}$ (of the order of one in a million). If this occurred, it would clearly indicate that the aspirin is effective.

The probability that the flower with the aspirin lasts longer in a minimum of 19 cases is around 2 in 10,000, whereas for a minimum of 15 cases it is around 2%, and for 14 cases is around 6%. As such, it is not so strange if the flower with aspirin lasts in 14 or more cases if it is ineffective. Taking the probability of error as 5% (referred to as the 'level of significance'), it should be considered that the aspirin is effective if it wins in a minimum of 15 cases.

Minimum no. of times flower with aspirin lasts longer	Probability of occurring if the aspirin is not effective
20	0.00000095
19	0.00002003
18	0.00020123
17	0.00128841
16	0.00590897
15	0.02069473
14	0.05765915

This is an extremely simple test and does not take into account whether one flower outlasts the other by a day or a week. There are other tests, such as the so-called Wilcoxon paired sample test which does account for the difference in each pair. But the most important factor is not the test which is chosen but ensuring that the experiment has been suitably designed and carried out, and that conclusions are not extrapolated beyond what has been proven.

Do expensive batteries last longer?

When we purchase a device for listening to music, we do not only make our selection based on its features but also because we prefer it over another. However, when we purchase batteries for it, the only important factor is how long they last.

Furthermore, it is interesting to observe the difference in prices between the same type of batteries, depending on the brand or the place where they are purchased. Normal 1.5-volt batteries can be twice the price if they are made by a well-known brand when compared to cheaper generic brands from supermarkets (which are not necessarily of a poor quality). It is also the case that there have recently been special offers for well-known brands and that the differences in price are no longer as great (the market imposes its laws).

Do expensive batteries last longer? And if this is the case, is it worth buying them? Or rather, does the increased duration compensate for the increase in price. To respond to these questions, we need data: we must carefully design a procedure to obtain them and then carry out a suitable analysis in order to reach our conclusions. Hence, we will need to use statistics.

HOW TO DIVIDE 20 RATS INTO TWO RANDOM GROUPS OF 10

Let's suppose we are carrying out a research project with laboratory rats to compare their stamina when they follow a certain kind of diet. Let's call this diet, which is "rich in saturated fatty acids" A and the other B. There are 20 similar rats which are about the same age and have the same general properties. They must be randomly divided into two groups of 10 and each group is fed on the corresponding diet.

After a number of months of being looked after, the rats are subjected to a test of their stamina which consists of making them swim in a pool and counting the time until they are no longer able to stay on the surface (at this point they are rescued).

The results show that the group that has followed diet B has more stamina than that which followed diet A (the average times of the different groups clearly show a significant difference in favour of B), and we are delighted with the result. But how were the rats separated? Randomly of course, by putting our hand in the cage and 'randomly' removing one after another until reaching 10. These rats made up group A and those which were left in the cage made up group B. Is there a problem here?

Of course. If the selection was carried out in this way, it was not random. Separating the rats in this way (putting our hand in and removing the first rat we can catch) will tend to catch the slowest rats, ones which are weaker or have slower reflexes (the others hide first), and these (group A) are the rats which are slower in the experiment. But are they slower as a result of their diet, or is it just that we have chosen the slowest in the group? There is no way of knowing. The moral of the story is that it is extremely important to ensure that the distribution of the groups to be compared is completely random, using numbers, paper or whatever is necessary. An error at this stage is difficult to fix.

The problem is not an easy one, as the following examples show:

1. The duration of batteries varies, both among expensive and cheap ones. They cannot be compared one by one because it is clear that the duration will be different (if we measure them with suitable precision), and hence even if a type lasts longer on one occasion this will not always be the case.

2. If we take a sample of batteries of each type and compare the average duration of each sample, the fact that one average is greater than the other still does

not provide any guarantees. If all the batteries were of the same brand and we divided them into two groups, we can be sure that the average of one group would be different to that of the other. The difference needs to be 'statistically significant'.

3. Batteries have different applications, and different discharge rhythms; while the duration for some applications may be the same, for others it may differ.

4. Measuring the duration of batteries is not easy. We cannot be present all day (and night!) waiting for a device to stop working.

It is possible to choose a type of application and compare the durations of a sample of expensive batteries and a sample of cheap ones. For a discharge rate similar to that of a torch, we can connect a battery to an alarm clock (with hands, digital ones are no use) and a light bulb (from a torch) as shown in the following figure: when the battery dies, the alarm clock will stop and by looking at the time at which it stopped working, we can know just how long the battery has taken to discharge. It will be necessary to check at least every 12 hours, but under these conditions, the batteries will not last long.

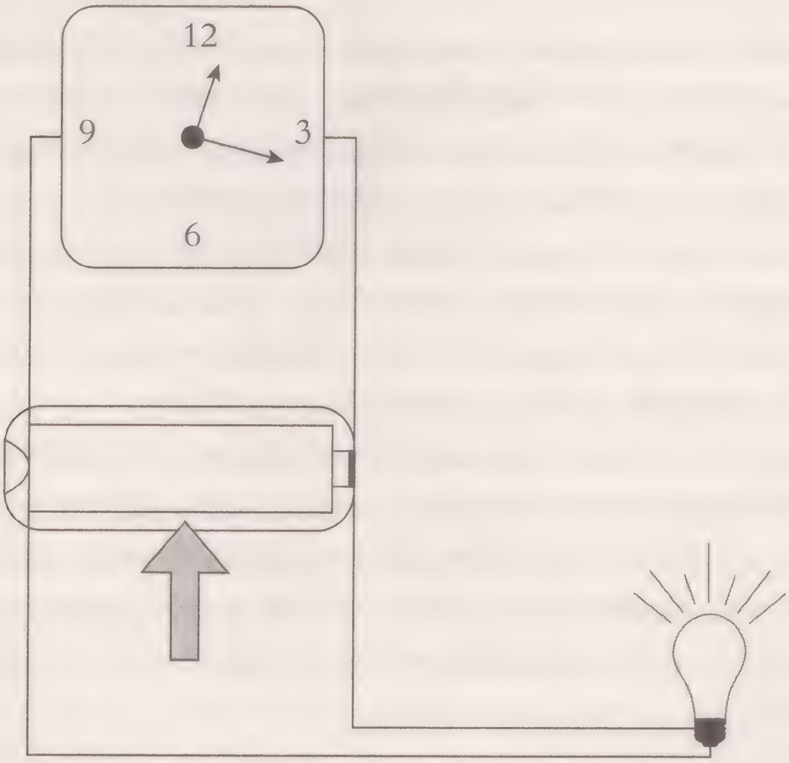
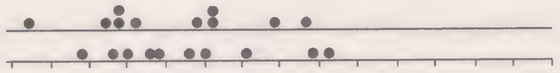
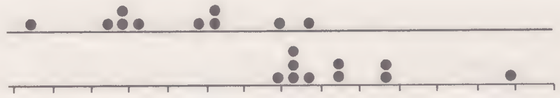
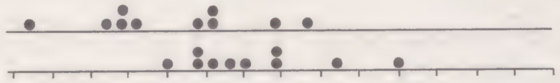


Diagram for measuring how long a battery lasts.

To analyse the data we have obtained, it is always a good idea to work with a graphical representation. In a case such as this one, with a small number of samples (e.g. 10 batteries of each type), it is enough to represent them using scatter graphs and compare them. It could be that there are no differences, that the differences are clear, or that the result is dubious. The statistical tests must confirm this initial impression: We cannot have the graphs saying one thing and the test telling us another.

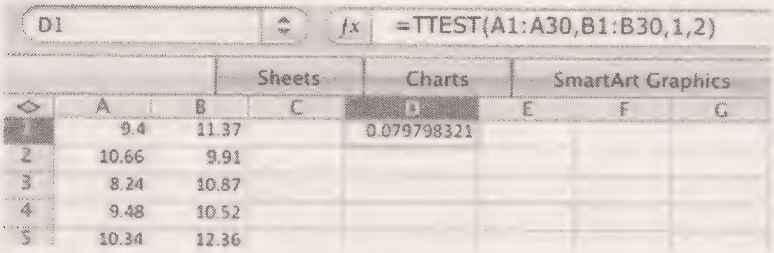
	Differences are not appraised
	The average of the second group is more than that of the first one.
	It is not clear that the difference is significant

Three possible situations analysed graphically.

To analyse the data obtained in this way, we can apply the Student t test for independent samples. This is easily done using a calculation tool such as Excel. All we need to do is indicate the position of the data, whether it is a test with one or two tails and which variant of the test we wish to analyse.

The part about one or two tails makes reference to the alternative hypothesis (the null hypothesis is that there is no difference). If the more expensive batteries last longer, as can be reasonably expected, the test has just one tail. If the alternative is that they have different durations, the test has two tails.

With respect to the type of test, it should be indicated whether it involves paired samples and, in the event that it does not, as in our example, if we can consider the variance to be equal in both situations. If the data has the same appearance as in the previous scatter graphs, there is no problem in specifying that the variance is the same. In the event of doubt, we can indicate that they are unequal; the result hardly changes.



=TTEST(A1:A30,B1:B30,1,2)

	A	B	C	D	E	F	G
1	9.4	11.37		0.079798321			
2	10.66	9.91					
3	8.24	10.87					
4	9.48	10.52					
5	10.34	12.36					

Using Excel to obtain the p-value in a Student t test.

In the dubious case that corresponds to the third scatter graph, the test tells us that the p-value is 0.08 (there is no point in including all the decimal places that are in Excel) and, as we already know, this means that if on average there is no difference in the duration between the different types of batteries, a difference as large as the one we have obtained randomly occurs 8% of the time.

Do bags of water keep flies away?

Hanging transparent plastic bags filled with water is a popular home remedy to keep away flies (on the Internet references ranging from Latin America to Thailand can be found). Some people believe it works, while others do not.

What is curious however, is that the people who believe it to work do so for a range of reasons: some argue that the light is broken down as it passes through the bag of water and this leaves the fly disorientated as a result of having compound eyes. Others claim that flies avoid coming close to the water because they know that if they get wet, they will not be able to fly. Finally, there are also those who believe that they are useful for precisely the opposite reason: they are hung in shops to attract flies so that they do not bother the public.

Do they work? Without going into the reasons, the only way of answering this question is to gather data through a well-designed experiment and see what conclusions can be drawn. It goes without saying that this will not be easy. The idea would be to count the number of flies in a space with and without bags. On some days, the bags would be left hanging and on others they would not (at random) and the number of flies would be counted every day.

However, counting flies is not easy, although technology may help us: certain cameras can be programmed to take a set of photographs at a given interval and perhaps a good shot on a white background would give photos that would allow us to count the number of flies at a given moment in time. In the event that this method was viable, there is the drawback that it is not possible to make out if the flies are different or if they change over time. Another way of indirectly counting how many flies are present in the area would be to use adhesive strips which catch the flies…

Doubtless the reader will be able to think of other ways. What is certain however, is that if the data is not gathered in a well-designed experiment, we will not know if the bags are useful.

Bibliography

BLASTLAND, M. and DILNOT, A., *The Tiger that isn't: Seeing through a World of Numbers*, London, Profile Books, 2008.

SALSBURG, D., *The Lady Tasting Tea. How Statistics Revolutionized Science in the 20th Century*, New York, W.H. Freeman, 2001.

TANUR, J.M. ET AL., *Statistics: A Guide to the Unknown*, San Francisco, Holden Day, 1972.

Index

Absolute Certainty and Other Fiction

The secrets of statistics

Statistics has been described as the practice of "torturing numbers to make them confess". This observation possibly has its roots in the fact that any examination of statistics starts from the conviction that 'certain' means little more than 'highly probable'. However, the study of statistics is undoubtedly the most important branch of applied mathematics and can often represent our best guide to taking crucial decisions in the face of uncertainty.